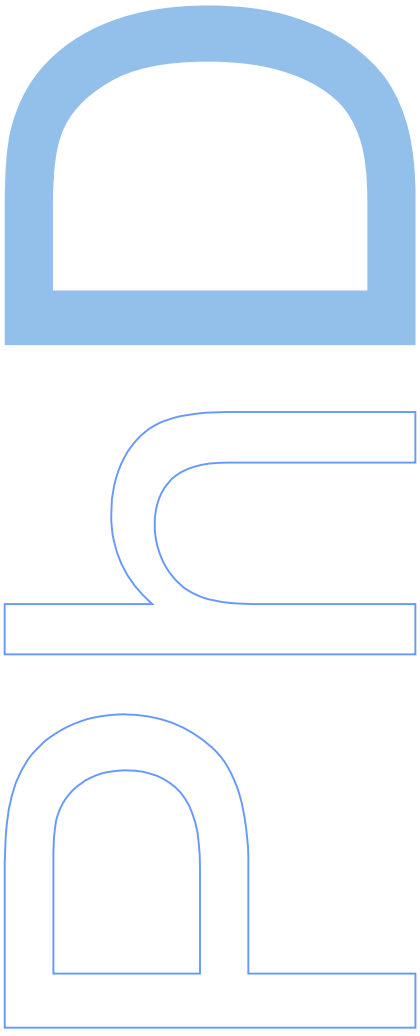
Filtragem de soluções espectrais via aproximação de Padé

João Emílio Raimundo Carrilho de Matos
Programa Doutoral em Matemática Aplicada
Departamento, de Matemática
2015

Orientador
Professor Doutor
José Manuel Andrade de Matos
Professor coordenador,
Instituto Superior de Engenharia do Porto

Coorientador
Professora Doutora
Maria João Pinto Sampaio Rodrigues
Professora Auxiliar,
Faculdade de Ciências da Universidade do Porto



à Carla
aos meus Pais

Agradecimentos

É com grande satisfação ter tido como orientadores científicos deste trabalho o Professor Doutor José Matos e a Professora Doutora Maria João Rodrigues. A ambos agradeço a disponibilidade, o interesse, a revisão deste trabalho e a amizade.

À Professora Doutora Ana Cristina Matos da Universidade de Lille, agradeço a disponibilidade pessoal e as sugestões que em muito contribuíram para a realização deste trabalho.

Ao Professor Doutor Bernhard Beckermann da Universidade de Lille, o meu agradecimento pela excelente hospitalidade com que me recebeu em Lille. Agradeço especialmente a sua amizade e as sugestões particularmente proliferas.

Ao Professor Doutor Manuel Cruz pelo incentivo e pelo facto de estar sempre disponível.

Ao Instituto Politécnico Português pela ajuda institucional pela concessão de dispensa de serviço ao abrigo do programa de apoio à formação avançada.

À Carla um agradecimento especial pelo apoio constante, em especial, nos momentos difíceis.

Resumo

Neste trabalho investiga-se o uso de aproximantes de Padé (AP) para filtrar soluções de equações diferenciais obtidas por métodos espectrais. A ideia chave de usar AP para filtrar soluções espectrais é aproveitar as propriedades dos AP, nomeadamente: acelerar a convergência das somas parciais de séries, permitem estimar singularidades usando para o efeito os pólos dos AP e alargam o domínio de convergência de séries.

No primeiro capítulo começamos por descrever resumidamente a teoria que envolve as diferentes formulações dos métodos espectrais e desenvolvemos novos procedimentos que permitem estabilizar numericamente os métodos espectrais. Deste modo alterámos o algoritmo usado por E. L. Ortiz e H. Samara no cálculo de soluções Tau de forma a evitar transformações de semelhança. Além disso modificámos os algoritmos de colocação-Galerkin proposto por J. P. Boyd, para equações diferenciais ordinárias e para equações em derivadas parciais, de modo a ser possível aplicar o método de filtragem referido neste trabalho.

O segundo capítulo é dedicado ao estudo dos AP de séries unidimensionais. Resumimos a teoria da AP de séries de potências e de séries de polinómios ortogonais. Tendo em conta as dificuldades inerentes ao cálculo de AP precisos, damos especial ênfase à componente numérica do cálculo destes aproximantes. Para superarmos parcialmente estas dificuldades, com AP de séries ortogonais, seguimos uma abordagem similar ao conceito de AP robustos de séries de potências introduzidos por P. Gonnet, S. Guttel e L. N. Trefethen. Tendo em vista estimar singularidades, no caso de AP de séries de Chebyshev (ACP) e de séries de Legendre (ALP), foram deduzidas relações que permitem o cálculo de pólos de AP do tipo $(p, 1)$, para AP de séries de famílias de polinómios ortogonais que satisfazem a condição de Szëgo, e do tipo $(p, 2)$, para ACP. O uso destas relações permite estimar singularidades sem ser necessário recorrer ao cálculo de AP.

No terceiro capítulo apresentamos os resultados obtidos na filtragem de soluções espectrais de equações diferenciais ordinárias rígidas (stiff). Nas conclusões estabelecemos algumas regras heurísticas a ter em conta na aplicação deste método de filtragem.

O quarto capítulo é dedicado à filtragem de solução espectrais de equações em derivadas parciais, em duas variáveis, via AP bidimensional. Começamos por descrever as dificuldades obtidas quando passamos da AP unidimensional para a AP multidimensi-

onal. Introduzimos as várias abordagens de AP multidimensional, tanto para séries de potências como séries de polinómios ortogonais. Implementámos o algoritmo sugerido por A.C. Matos para o cálculo ACP bidimensionais *nested* e fornecemos algoritmos que permitem o cálculo de alguns ACP bidimensionais provenientes de equações definidas por reticulados. Nomeadamente os ACP bidimensionais mistos e os ACP bidimensionais "homogéneos" do tipo I e do tipo II. Finalmente apresentamos vários exemplos onde usámos estes ACP para filtrar soluções espectrais da equação de Poisson a duas variáveis.

Ao longo deste trabalho iremos usar, por uma questão de simplicidade, indistintamente as abreviaturas AP, ACP e ALP para referir aproximante ou aproximação (e respectivos plurais) de Padé, de Chebyshev Padé e de Legendre Padé respetivamente. Pensamos que o significado destas abreviaturas ficará claro no contexto em que são usadas.

Abstract

In this work we study the use of Padé approximants (PA) to filter solutions of differential equations obtained by spectral methods. The key idea to use PA as filters of spectral solutions is to take advantage of the properties of PA. More specifically, PA increase the rate of convergence of the series' partial sums, the poles of PA are good estimates of singularities and PA extend the domain of convergence of the series.

In the first chapter we start describing the theory involved in the different formulation of spectral methods and we develop new procedures that allow us to numerically stabilize the spectral methods. To this end we have modified the algorithm of the Tau method, proposed by E. L. Ortiz and H. Samara, in order to avoid the use of similarity transformations. Furthermore we have also modified the scheme of collocation-Galerkin, proposed by J. P. Boyd, in order to make possible for it to be applied in our filtering.

The second chapter, is dedicated to the study of PA of unidimensional series. Here we outline the theory of PA from power series and from orthogonal polynomials' series. In view of the difficulty of the numerical computation of precise PA, we give special emphasis to the numerical computation of these PA. In order to overcome, partially, the difficulties with the computation of PA of orthogonal polynomials' series we follow a similar approach to the concept of robust PA of power series introduced by P. Gonnet, S. Guttel and L. N. Trefethen. Having in mind the estimation of singularities of functions given by Chebyshev (ACP) and Legendre (ALP) series, we have deduced some relations that allow the computation of poles of PA of type $(p, 1)$ for families orthogonal polynomials that satisfy the Szëgo condition as well as the computation of poles of PA of type $(p, 2)$ for ACP. The use of these relations allows the estimation of singularities without the actual computation of the PA.

In the third chapter we present the results obtained by applying the filtering method to stiff ordinary differential equations. In the last section we give some heuristic rules to take into account when applying this filtering method.

The fourth chapter is dedicated to the filtering of spectral solutions of partial differential equations in two variables using bidimensional PA. We start by describing the difficulties when we move from unidimensional PA to multidimensional PA. We introduce several approaches to multidimensional PA, both for power series and for orthogonal

polynomials. We have implemented the algorithm suggested by A. C. Matos for the computation of bidimensional nested ACP and we provide algorithms for the computation of some bidimensional ACP arising form equations defined by lattices, namely for mixed bidimensional ACP as well as for homogeneous bidimensional ACP of types I and of type II. Finally we exhibit several examples where we used these ACP to filter spectral solutions of Poisson's equation in two variables.

Résumé

Dans ce manuscrit on s'intéresse à l'utilisation des Approximants de Padé (AP) pour filtrer les solutions d'équations différentielles obtenues par des méthodes spectrales. L'idée clé dans l'utilisation des AP pour filtrer les solutions spectrales consiste à profiter des propriétés des AP, notamment le fait que les AP accélèrent la convergence des sommes partielles de séries, permettent d'estimer des singularités en utilisant les pôles des AP et élargissent le domaine de convergence des séries. Dans le premier chapitre nous commençons par décrire d'une forme résumée la théorie des différentes formulations des méthodes spectrales et nous développons des procédés qui permettent de stabiliser numériquement les méthodes spectrales. On change ainsi l'algorithme utilisé par E.L. Ortiz et H. Samara pour le calcul de solutions Tau de façon à éviter les transformations d'équivalence. Nous avons aussi modifié les algorithmes de collocation Galerkin proposés par J.P. Boyd, aussi bien pour les équations différentielles ordinaires que pour les équations aux dérivées partielles, de façon à ce que l'on puisse appliquer la méthode de filtrage développée dans ce travail. Le deuxième chapitre est dédié à l'étude des AP de séries unidimensionnelles. On résume la théorie des AP pour des séries de puissances et des séries de polynômes orthogonaux. Étant donné les difficultés dans le calcul des AP avec précision élevée, on s'intéresse spécialement à la composante numérique du calcul de ces approximants. Pour dépasser partiellement ces difficultés, avec les AP pour séries orthogonales, on suit une méthode similaire au concept de AP robustes introduits pour les séries orthogonales par P. Gonnet, S. Guttel et L.N. Trefethen. Dans le but d'estimer des singularités dans le cas de séries de Tchebyshev (ACP) et de séries de Legendre (ALP), nous avons déduit des relations qui permettent le calcul des pôles de AP de tipe $(p, 1)$ pour des AP de séries de familles de polynômes orthogonaux qui vérifient la condition de Szëgo, et de type $(p, 2)$ pour ACP. L'utilisation de ces formules permet d'estimer des singularités sans qu'il soit nécessaire de recourir au calcul des AP Dans le troisième chapitre nous présentons les résultats obtenus dans le filtrage de solutions spectrales d'équations différentielles ordinaires avec rigidité (stiff). Dans les conclusions nous établissons quelques règles heuristiques qui doivent être vérifiées dans l'application de cette méthode. Le quatrième chapitre est dédié au filtrage de solutions spectrales d'équations aux dérivées partielles à deux variables, via AP bidimensionnels. Nous commençons par décrire les

difficultés obtenues quand on passe de l'approximation unidimensionnelle à multidimensionnelle. Nous introduisons les différentes approches de AP multidimensionnelle, aussi bien pour les séries de puissances que pour les séries de polynômes orthogonaux. Nous avons implémenté l'algorithme suggéré par A. C. Matos pour le calcul ACP bidimensionnels « nested » et nous avons fourni des algorithmes qui permettent le calcul de quelques ACP bidimensionnels provenant d'équations définies par des réticulés. Notamment les ACP bidimensionnels mixtes et les ACP bidimensionnels homogènes de type I et de type II. Ensuite nous présentons différents exemples où nous avons utilisé ces ACP pour filtrer des solutions spectrales de l'équation de Poisson à deux variables. Tout le long de ce travail nous utiliserons, pour une question de simplicité, les abréviations AP, ACP, et ALP pour représenter approximation ou approximant de Padé, Chebyshev-Padé ou Legendre-Padé respectivement. La signification deviendra claire dans le contexte où c'est utilisée.

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Métodos Espectrais

Os métodos espectrais permitem encontrar aproximantes de soluções de equações diferenciais. As componentes chave na formulação de um método espectral são: a escolha das funções base e a escolha das funções teste (também conhecidas por funções peso). As funções tentativa são combinações lineares de funções base e representam uma aproximação da solução exata da equação diferencial. As funções teste são usadas para assegurar que a função tentativa satisfaça a equação diferencial, e eventualmente as condições suplementares, da forma mais exata possível. Traduz-se esta condição por uma forma de minimização do resíduo relativamente a uma norma adequada, resultante de usarmos expansões truncadas em vez da solução exata. Por esta razão os métodos espectrais podem ser vistos como casos especiais dos métodos dos resíduos ponderados [FS66]. Por outro lado, os métodos espectrais são um caso especial dos métodos de Galerkin-Petrov [ZC67], dado que o resíduo satisfaz uma condição de ortogonalidade relativamente às funções peso.

Existem essencialmente três abordagens na construção de um método espectral, nomeadamente, método de Galerkin, método de colocação e método Tau. Na abordagem de Galerkin as funções teste são iguais às funções tentativa e satisfazem individualmente algumas ou todas as condições suplementares. Exigimos, nesta abordagem, que o integral do produto do resíduo com cada função teste seja nulo. Na abordagem de colocação as funções teste são translações da função delta de Dirac centradas em certos pontos, chamados pontos de colocação. Na abordagem do método Tau, as funções teste não necessitam de satisfazer individualmente as condições suplementares pelo que é necessário acrescentar um conjunto de equações por forma a que as aproximações satisfaçam as condições suplementares.

1.1 Formulação de aproximações espectrais de problemas lineares estacionários

Seja Ω um subconjunto aberto e limitado de $\mathbb{R}^n$ com fronteira $\partial\Omega$ de classe C^∞ aos pedaços. Pretendemos encontrar uma aproximação da solução do problema com condições fronteira

$$\mathscr{L}u = f, \quad \text{em } \Omega, \tag{1.1}$$

$$\mathscr{B}u = 0, \quad \text{em } \partial\Omega_b, \tag{1.2}$$

onde $\mathscr{L}$ é um operador diferencial linear, e $\mathscr{B}$ é um conjunto de funcionais lineares que traduzem as condições fronteira do problema definidos num conjunto $\partial\Omega_b$ onde, $\partial\Omega_b \subset \partial\Omega$. Assumimos que existe um espaço de Hilbert $\mathbf{X}$ tal que $\mathscr{L}$ é um operador não limitado em $\mathbf{X}$. Representamos por $(*,*)$ o produto interno definido em $\mathbf{X}$, e a sua norma associada por $||*||$. Frequentemente tem-se $\mathbf{X} = L^2_w(\Omega)$, onde $L^2_w(\Omega)$ representa o espaço das funções mensuráveis tais que $\int_\Omega |f|^2 w < \infty$ e w é a uma função peso adequada (B.2). O domínio de $\mathscr{L}$ é $D(\mathscr{L}) \subset \mathbf{X}$ definido por

$$D(\mathscr{L}) = \{f \in \mathbf{X} \mid \mathscr{L}f \in \mathbf{X}\}$$

É suposto igualmente que $D(\mathscr{L})$ é um subespaço denso em $\mathbf{X}$.

Exemplo 1.1.1. [CHQZ07] Considerando o operador $\mathscr{L} = -\dfrac{d^2}{dx^2}$, no intervalo $\Omega =]-1,1[$. Tomando para função w a função peso de Legendre $w(x) = 1$ ou a função peso de Chebyshev $w(x) = (1-x^2)^{-\frac{1}{2}}$, então $\mathbf{X} = L^2_w(]-1,1[) = \{f :]-1,1[\to \mathbb{R} \mid (f,f) < \infty\}$, com $(f,g) = \int_{-1}^{1} fgwdx$. Então $\mathscr{L}$ é um operador não limitado cujo domínio é

$$D(\mathscr{L}) = \left\{ f \in C^1(]-1,1[) \mid \frac{d^2 f}{dx^2} \in L^2_w(]-1,1[) \right\},$$

onde se considera a derivação fraca (B.3).

Assumimos que os operadores funcionais que representam as condições fronteira fazem sentido quando aplicados a todas as funções de $D(\mathscr{L})$. Restringimos o domínio de $\mathscr{L}$ ao subespaço $D_B(\mathscr{L})$ de $D(\mathscr{L})$ definido por

$$D_B(\mathscr{L}) = \{f \in D(\mathscr{L}) \mid \mathscr{B}f = 0 \text{ em } \partial\Omega_b\},$$

que assumimos ser igualmente denso em $\mathbf{X}$. Logo consideramos o operador $\mathscr{L}$ a actuar entre $D_B(\mathscr{L})$ e $\mathbf{X}$

$$\mathscr{L} : D_B(\mathscr{L}) \subset \mathbf{X} \longrightarrow \mathbf{X},$$

e podemos escrever o problema (1.1), (1.2) da forma

$$\begin{aligned} &\mathscr{L}u = f, \\ &u \in D_B(\mathscr{L}), \end{aligned} \tag{1.3}$$

para $f \in \mathbf{X}$, onde a igualdade é entre duas funções (equivalentes) em $\mathbf{X}$.

No exemplo anterior o operador $\mathscr{L}$ pode ser complementado, por exemplo, ou com uma condição fronteira de Dirichlet, $\mathscr{B}u \equiv u(\pm 1) = 0$, ou com condições fronteira de Neumann, $\mathscr{B}u \equiv u'(\pm 1) = 0$. Em ambos os casos as condições fronteira fazem sentido, dado que as funções em $D_B(\mathscr{L})$ têm primeira derivada contínua. A densidade de $D_B(\mathscr{L})$ em $L^2_w(]-1,1[)$ é uma consequência da densidade de $C^\infty(]-1,1[)$ em $L^2_w(]-1,1[)$ [CHQZ07].

A primeira equação em (1.3) pode escrever-se da forma equivalente

$$(\mathscr{L}u, v) = (f, v), \quad \forall v \in \mathbf{X}. \tag{1.4}$$

O primeiro membro (1.4) é uma forma bilinear em $D_B(\mathscr{L}) \times \mathbf{X}$ e será representada por $a(u, v)$. De forma semelhante o segundo membro de (1.4) é uma forma linear em $\mathbf{X}$ e será representada por $F(v)$. Então o problema (1.3) pode escrever-se na forma

$$\begin{aligned} &a(u, v) = F(v), \quad \forall v \in \mathbf{X}, \\ &u \in D_B(\mathscr{L}). \end{aligned} \tag{1.5}$$

A forma bilinear pode ser definida por uma expressão equivalente definida num espaço $W \times V$ que é mais apropriado para mostrar que o problema (1.3) é bem definido e para definir uma aproximação numérica. O espaço W contém funções menos regulares que as funções de $D_B(\mathscr{L})$, enquanto o espaço V contém funções mais regulares do que as funções de $\mathbf{X}$. Geralmente a expressão de $a(u, v)$ é obtida aplicando integração por partes e usando as condições fronteira. Como exemplo, considerando a função peso de Legendre no exemplo 1.1.1, tem-se

$$(\mathscr{L}u, v) = \int_{-1}^{1} -\frac{d^2u}{dx^2} v dx = \int_{-1}^{1} \frac{du}{dx}\frac{dv}{dx} = a(u, v),$$

desde que $\dfrac{dv}{dx} \in L^2(]-1,1[)$ e que em ambos os extremos do intervalo $[-1, 1]$, uma das funções $\frac{du}{dx}$ ou v se anule.

Introduzidos os espaços W e V, a formulação do problema (1.5) fica com a forma

$$\begin{aligned} &a(u, v) = F(v), \quad \forall v \in V, \\ &u \in W. \end{aligned} \tag{1.6}$$

Evidentemente, será conveniente verificar se um dado problema da forma (1.1)-(1.2) está, ou não, bem definido. Para garantir que um dado problema está bem definido, suponhamos que existe um espaço de Hilbert $E \subset \mathbf{X}$ denso em $\mathbf{X}$ com norma $||*||_E$, para o qual existe uma constante positiva C tal que $||u|| \leq C||u||_E$, para todo $u \in E$ e que $D_B(\mathscr{L}) \subset E$ é denso em E. Assumimos que a forma bilinear $a(u, v)$ está definida em $E \times E$ e que existem constantes $\alpha, A > 0$ tal que

$$\alpha||u||^2_E \leq a(u, u) \quad \forall u \in E \tag{1.7}$$

$$|a(u,v)| \leq A||u||_E||v||_E \quad \forall u, v \in E \tag{1.8}$$

Então os espaços W e V coincidem com E. A desigualdade (1.7) é a condição de coercividade para a forma bilinear $a(u,v)$ e estabelece que o operador $\mathscr{L}$, com as condições fronteira, é um operador positivo, que é coercivo sobre E. Por outro lado, (1.8) é uma condição de continuidade para $\mathscr{L}$, no sentido que na norma $||*||_E$ depende continuamente de u e de v. Como a forma linear F satisfaz a desigualdade $|F(v)| \leq \|f\| \cdot \|v\| \leq C\,\|f\| \cdot \|v\|_E$ concluimos que existe uma constante $C_F > 0$ tal que

$$|F(v)| \leq C_F||v||_E, \quad \forall v \in E \tag{1.9}$$

Sob as condições (1.7)-(1.9), o teorema de Lax-Milgram, [Bre83], assegura que existe uma solução única u do problema

$$\begin{aligned} &a(u,v) = F(v), \quad \forall v \in E, \\ &u \in E. \end{aligned} \tag{1.10}$$

Além disso, u depende continuamente de f, uma vez que se tem

$$||u||_E \leq \frac{C}{\alpha}||f||,$$

e é solução do problema (1.3).

Voltando de novo ao exemplo 1.1.1 com as condições de Dirichlet, as condições (1.7)-(1.9) são satisfeitas com $E = H^1_{0,w}$ (B.22) que é um espaço de Hilbert com norma

$$||u||_E = \left(\int_{-1}^{1} (u')^2\, w dx\right)^{\frac{1}{2}}.$$

O resultado é imediato para as funções peso de Chebyshev e de Legendre [CHQZ07]. Note-se que todas as funções em E satisfazem as condições fronteira. Por outro lado se tivermos o operador $\mathscr{L} = -\dfrac{d^2}{dx^2} + I$, as condições de Neumann e $w(x) = 1$, então as condições (1.7)-(1.9) são igualmente satisfeitas com $E = H^1(]-1,1[)$ (B.4) que é um espaço de Hilbert para a norma

$$||u||_E = \left(\int_{-1}^{1} (u^2 + (u')^2) dx\right)^{\frac{1}{2}}.$$

Neste caso, as funções em E não satisfazem necessariamente as condições fronteira, mas a solução de (1.10) satisfaz.

A condição de positividade, (1.7), verifica-se quase imediatamente para o problema (1.3). Contudo esta não é a situação geral. Em muitos problemas podemos usar uma condição mais geral, conhecida por *condição de inf-sup* , [CHQZ07], que apresentaremos de seguida.

Sejam $W \subset \mathbf{X}$ e $V \subset \mathbf{X}$ espaços de Hilbert, com normas $||*||_W$ e $||*||_V$, respectivamente. Assumimos que a inclusão de V em $\mathbf{X}$ é contínua, i.e. existe uma constante $C > 0$ tal que $||v|| \leq C||v||_V \ \forall v \in \mathbf{X}$. Supomos que $D_B(\mathscr{L})$ está contido densamente em W e que V está contido densamente em $\mathbf{X}$. Assumimos ainda que a forma bilinear $a(u, v)$ está definida em $W \times V$, e que existem constantes $\alpha > 0$ e $A > 0$ tais que

$$0 < \sup_{u \in W} a(u, v), \quad \forall v \in V \setminus \{0\}, \tag{1.11}$$

$$\alpha||u||_W \leq \sup_{v \in V \setminus \{0\}} \frac{a(u, v)}{||v||_V}, \quad \forall u \in W, \tag{1.12}$$

$$|a(u, v)| \leq A||u||_W||v||_V, \quad \forall u \in W \ \text{ e } \ \forall v \in V. \tag{1.13}$$

Usando uma versão do teorema de Lax-Milgram [Neč62], as condições (1.11)-(1.13) em conjunto com (1.9) asseguram que o problema (1.5) tem uma solução única que depende continuamente dos dados, i.e.

$$||u||_W \leq \frac{C}{\alpha}||f||.$$

Escolhendo $V = W = E$, a condição de coercividade (1.7) implica as condições (1.11) e (1.12) [CHQZ07].

1.2 Aproximações Espectrais

Optando por uma das formulações (1.5) ou (1.10) do problema (1.3), pode-se definir, sem especificar ainda o método espectral, uma aproximação espectral do problema (1.3) da seguinte forma

$$\begin{aligned} &a_N(u_N, v) = F_N(v), \quad \forall v \in Y_N, \\ &u_N \in X_N, \end{aligned}$$

onde N é um inteiro positivo, X_N e Y_N são subespaços de dimensão finita com a mesma dimensão, a_N é a restrição da forma bilinear a a $X_N \times Y_N$ e F_N a restrição da forma linear F a Y_N. Dependendo de como as condições suplementares são impostas, X_N pode estar contido em $D_B(\mathscr{L})$ (caso cada função de X_N satisfaça as condições suplementares) ou não estar contido em $D_B(\mathscr{L})$ (neste caso a solução espectral apenas satisfaz as condições suplementares aproximadamente).

No *método de Galerkin* restringimos os espaços das funções tentativa e das funções teste em (1.10) ou (1.5) ao espaço de dimensão finita X_N, ou seja, o método de Galerkin para o problema (1.10) toma a forma

$$\begin{aligned} &a(u_N, v) = F(v), \quad \forall v \in X_N, \\ &u_N \in X_N \end{aligned} \tag{1.14}$$

O *método de Galerkin com integração numérica* (G-IN) é obtido da formulação (1.14) pela substituição dos integrais que definem a forma bilinear a, e a forma linear F, por fórmulas de quadratura. As formas resultantes dependem da ordem N, da aproximação espectral e são representadas com o índice N. Um método de G-IN pode ser representado na forma

$$\begin{aligned} &a_N(u_N, v) = F_N(v), \quad \forall v \in X_N, \\ &u_N \in X_N. \end{aligned} \tag{1.15}$$

As formulações (1.14) ou (1.15) são bastante gerais, dado que têm em conta possíveis condições fronteira impostas no sentido fraco. Uma formulação mais restrita para os métodos espectrais é suficiente quando os aproximantes são de classe $C^\infty(\Omega)$ (geralmente são polinómios), logo o operador $\mathscr{L}$ está definido em X_N. Quando as funções tentativa satisfazem individualmente as condições fronteira, i.e., quando $X_N \subset D_B(\mathscr{L})$, a forma bilinear pode-se escrever na forma forte $a(u, v) = (\mathscr{L}u, v)$ para todo $u \in X_N$. Então o método de Galerkin fica da forma

$$\begin{aligned} &(\mathscr{L}u, v) = (f, v), \quad \forall v \in X_N, \\ &u_N \in X_N. \end{aligned} \tag{1.16}$$

O *método Tau* é obtido permitindo que as funções teste estejam num espaço Y_N diferente de X_N. O espaço Y_N tem a mesma dimensão de X_N, mas as funções de Y_N não necessitam de satisfazer individualmente as condições fronteira, como as funções de X_N. A formulação do método Tau é escrita da forma

$$\begin{aligned} &(\mathscr{L}u_N, v) = (f, v), \quad \forall v \in Y_N, \\ &u_N \in X_N. \end{aligned} \tag{1.17}$$

O *método de colocação* é escrito de uma forma idêntica a (1.16)

$$\begin{aligned} &(\mathscr{L}_N u, v)_N = (f, v)_N, \quad \forall v \in X_N, \\ &u_N \in X_N. \end{aligned} \tag{1.18}$$

onde, $\mathscr{L}_N$ é uma aproximação de $\mathscr{L}$, obtida frequentemente por substituição das derivadas exatas pelas derivadas de interpolação [CHQZ07] e a forma bilinear $(u, v)_N$ é, frequentemente, definida pelos valores que as funções u e v tomam nos pontos de colocação. Esta forma é definida num subespaço $Z \subset X$ cujos elementos são funções contínuas para as quais o valor pontual tem significado e assumimos que $\mathscr{L}_N$ envia elementos de X_N em elementos de Z e que $f \in Z$.

Podemos juntar as formulações (1.16), (1.17) e (1.18) na forma

$$\begin{aligned} &(\mathscr{L}_N u - f, v)_N = 0, \quad \forall v \in Y_N, \\ &u_N \in X_N, \end{aligned} \tag{1.19}$$

para escolhas adequadas de $\mathscr{L}_N$, Y_N e $(*,*)_N$ que dependem do método aplicado. Esta forma mostra que os métodos espectrais pertencem à classe dos métodos dos resíduos ponderados. A escolha do espaço Y_N e do produto interno $(u,v)_N$ definido em Y_N define o modo em que minimizamos o resíduo.

Uma forma operacional equivalente a (1.19) é

$$\begin{aligned} &u_N \in X_N, \\ &Q_N \mathscr{L}_N u_N = Q_N f, \end{aligned} \tag{1.20}$$

onde $Q_N : Z \subset X \longrightarrow Y_N$, satisfaz

$$(z - Q_N z, v)_N = 0, \quad \forall v \in Y_N, \tag{1.21}$$

ou seja Q_N é a projecção ortogonal sobre o espaço Y_N no produto interno $(u,v)_N$.

Nas secções seguintes iremos resumir os resultados gerais que garantem a estabilidade e convergência dos métodos espectrais clássicos (métodos de Galerkin, de colocação, G-IN e Tau). Para métodos espectrais não clássicos, geralmente métodos que combinam dois (ou mais) métodos espectrais clássicos, o estudo da estabilidade e convergência pode ser feito usando os resultados para os quatro métodos clássicos.

Os elementos chave para cada método espectral são: o espaço das funções tentativa, o espaço das funções teste, a forma bilinear $a_N(u,v)$ e a forma linear $F_N(v)$. Na formulação operacional (1.20) consideramos que os elementos chave são: o operador projeção Q_N e o produto interno $(u,v)_w$. Tendo em vista formalizar os métodos espectrais mais relevantes iremos considerar que o domínio Ω do problema (1.1)-(1.2) é da forma

$$\Omega = \prod_{k=1}^{d} I_k,$$

onde $I_k =]0, 2\pi[$ ou $I_k =]-1, 1[$. Dado um inteiro positivo N a solução espectral do problema (1.1)-(1.2) é, frequentemente, uma combinação linear de funções que são polinómios, trigonométricos ou algébricos, de grau não superior a N. O espaço definido por todas estas combinações lineares com coeficientes em $\mathbb{R}$ é representado por $\mathcal{P}_N(\Omega)$. Se $d > 1$, N representa um vector de índices $N = (N_1, \cdots, N_d)$ e assumiremos sempre que $\mathcal{P}_N(\Omega) \subset D(\mathscr{L})$.

1.3 Método de Galerkin

Nos métodos de Galerkin as funções tentativa e as funções teste satisfazem as condições fronteira. Seja X_N o subespaço de $\mathcal{P}_N(\Omega)$ das funções que satisfazem as condições fronteira

tal que $X_N \subset D_B(\mathscr{L})$ e seja $\{\phi_k \mid k \in J\}$ uma base em X_N (não necessariamente ortogonal relativamente ao produto interno definido em $\mathbf{X}$), onde J é um conjunto de índices.

Um método de Galerkin é caracterizado pelas equações

$$\begin{aligned} &(\mathscr{L}u_N, \phi_k) = (f, \phi_k), \quad \forall k \in J, \\ &u_N \in X_N. \end{aligned} \tag{1.22}$$

As incógnitas do problema são os coeficientes $\hat{c}_k$, $k \in J$, da solução de Galerkin $u_N = \sum_{k \in J} \hat{c}_k \phi_k$. Equivalentemente pode-se escrever as equações (1.22) na forma

$$\begin{aligned} &(\mathscr{L}u_N, v) = (f, v), \quad \forall v \in X_N \\ &u_N \in X_N. \end{aligned} \tag{1.23}$$

Relativamente à formulação geral (1.19), os métodos de Galerkin são caracterizados pelas escolhas $Y_N = X_N$, $(u, v)_N = (u, v)$ (ou seja opta-se pelo produto interno definido em $\mathbf{X}$) e $\mathscr{L}_N = \mathscr{L}$.

Uma generalização dos métodos de Galerkin são os métodos de *Petrov-Galerkin* [CHQZ07] nos quais as funções testes e as funções tentativas são diferentes mas satisfazem igualmente as condições fronteiras. Nos métodos de Petrov-Galerkin tem-se $Y_N \neq X_N$ e as equações (1.23) tomam a forma

$$\begin{aligned} &(\mathscr{L}u_N, v) = (f, v), \quad \forall v \in Y_N \\ &u_N \in X_N. \end{aligned} \tag{1.24}$$

Estabilidade e Convergência

Uma condição suficiente para a estabilidade e convergência de uma aproximação de Galerkin é que a forma bilinear $a(u, v) = (\mathscr{L}u, v)$ satisfaça a condição de coercividade (1.7) e a condição de continuidade (1.8) e $X_N \subset E$, para todo X_N [CHQZ07]. Ou seja tem-se

$$\alpha ||u||_E^2 \leq (\mathscr{L}u, u), \quad \forall u \in X_N \tag{1.25}$$

e

$$|(\mathscr{L}u, u)| \leq A||u||_E||v||_E, \quad \forall u, v \in X_N. \tag{1.26}$$

1.4 Método de Colocação

No método de colocação usa-se um conjunto de pontos $x_k \in \overline{\Omega}$, $k \in J$, onde J é um conjunto de índices e o número de pontos usados deverá ser igual à dimensão do espaço $\mathcal{P}_N(\Omega)$. As condições fronteira serão impostas usando alguns pontos $x_k \in \partial\Omega$. Iremos

assumir que o conjunto J é tal que para todo $k \in J$ existe um único polinómio $\phi_k \in \mathcal{P}_N(\Omega)$ tal que

$$\phi_k(x_m) = \begin{cases} 1 & \text{se } k = m. \\ 0 & \text{se } k \neq m. \end{cases} \tag{1.27}$$

As funções ϕ_k são os polinómios característicos de Lagrange (D.2), para o caso unidimensional, e formam uma base para $\mathcal{P}_N(\Omega)$ e tem-se $v(x) = \sum_{k\in J} v(x_k)\phi_k(x)$ para todo $v(x) \in \mathcal{P}_N(\Omega)$.

Um método de colocação obtém-se dividindo o conjunto J em dois subconjuntos disjuntos J_e e J_b de forma a que $\{x_k \,|k \in J_b\,\} \subset \partial\Omega_b$. Seja $\mathscr{L}_N$ uma aproximação do operador $\mathscr{L}$ no qual as derivadas são aproximadas por interpolação nos pontos x_k, $k \in J$. A aproximação de colocação é um polinómio $u_N \in \mathcal{P}_N(\Omega)$ que satisfaz as equações

$$\mathscr{L}_N u_N(x_k) = f(x_k), \quad \forall k \in J_e, \tag{1.28}$$

$$\mathscr{B} u_N(x_k) = 0, \qquad \forall k \in J_b. \tag{1.29}$$

Nos métodos de colocação as incógnitas são os valores de $u_k \equiv u_k\,(x_k)$, $k \in J$, ou seja, os coeficientes de u_N relativamente à base de Lagrange (1.27). O conjunto J_b poderá ser eventualmente o conjunto vazio. Nestes casos as condições fronteira surgem implicitamente na definição do operador $\mathscr{L}_N$ ou introduzindo penalidades [CHQZ07].

Para inserir o esquema de colocação (1.28)-(1.29) no enquadramento da secção (1.1), fixa-se uma família de pesos $w_k > 0$ e introduz-se a forma bilinear $(u, v)_N$ no espaço $Z = C^0(\overline{\Omega})$ definida por

$$(u, v)_N = \sum_{k\in J} u(x_k)\overline{v(x_k)}w_k. \tag{1.30}$$

A existência da base de Lagrange (1.27) assegura que a forma bilinear (1.30) é um produto interno em $\mathcal{P}_N(\Omega)$. A base (1.27) é ortogonal relativamente a este produto interno. Deste modo podemos definir a *norma discreta* induzida em $\mathcal{P}_N(\Omega)$

$$\|u\|_N = \sqrt{(u, u)_N}, \quad \forall u \in \mathcal{P}_N(\Omega) \tag{1.31}$$

Para que este produto interno (definido em $\mathcal{P}_N(\Omega)$) aproxime *suficientemente* o produto interno $(*, *)$ definido em $\mathbf{X}$, assumiremos que os nós x_k e os pesos w_k verificam a condição

$$(u, v)_N = (u, v), \quad \forall u, v \text{ tais que } uv \in \mathcal{P}_{2N-1}(\Omega). \tag{1.32}$$

Em todas as aplicações se escolhermos os x_k como sendo nós das fórmulas de quadratura gaussiana (D.3), a condição (1.32) verifica-se.

Seja X_N o espaço dos polinómios definido por

$$X_N = \{v \in \mathcal{P}_N(\Omega) \mid \mathscr{B}v(x_k) = 0, \quad \forall k \in J_b\}\,. \tag{1.33}$$

O método de colocação pode escrever-se na forma

$$\begin{aligned} &u_N \in X_N, \\ &(\mathscr{L}_N u_N, \phi_k)_N = (f, \phi_k), \quad \forall k \in J_e. \end{aligned} \tag{1.34}$$

Definindo Y_N o espaço gerado por $\phi_k \in J_e$,

$$Y_N = \{v \in \mathcal{P}_N(\Omega) \mid v(x_k) = 0, \quad \forall k \in J_b\}, \tag{1.35}$$

então (1.34) fica na forma

$$\begin{aligned} &(\mathscr{L}_N u_N, v)_N = (f, v), \quad \forall v \in Y_N \\ &u_N \in X_N. \end{aligned} \tag{1.36}$$

Equivalentemente, podemos escrever (1.34) na forma operacional (1.21)

$$Q_N(\mathscr{L}_N u_N - f) = 0. \tag{1.37}$$

Nos métodos de colocação, $Q_N v$ é o polinómio de grau N tal que $Q_N v(x_k) = v(x_k)$ para todo $x_k \in J_e$ e $Q_N v(x_k) = 0$ para todo $x_k \in J_b$.

No caso de se ter condições de Dirichlet, i.e. $\mathscr{B}v = v$, então $X_N = Y_N$ e o método de colocação pode igualmente ser interpretado como um método de Galerkin com integração.

Nos seguintes exemplos seguimos o método da recombinação de bases indicado em [Boy01].

Exemplo 1.4.1. Pretende-se encontrar aproximações, usando um método de colocação na base de Chebyshev, de funções u definidas por equações diferenciais lineares de segunda ordem ordinárias com condições fronteira de Dirichlet. Ou seja, a função u é solução do problema

$$\mathscr{L}u = f, \quad \text{em } \Omega =]-1, 1[\tag{1.38a}$$

$$u(-1) = \alpha, \quad u(1) = \beta. \tag{1.38b}$$

onde

$$\mathscr{L}u := a(x)\frac{d^2u}{dx^2} + b(x)\frac{du}{dx} + c(x)u(x).$$

O problema (1.38a)-(1.38b) pode reescrever-se num problema equivalente com condições de Dirichlet homogéneas. Ou seja, resolve-se o problema

$$\mathscr{L}v = g, \quad \text{em } \Omega =]-1, 1[\tag{1.39a}$$

$$v(-1) = 0, \quad v(1) = 0. \tag{1.39b}$$

onde $v(x) = u(x) - B(x)$, $g(x) = f(x) - \mathscr{L}B(x)$ e $B(x) = \frac{\alpha}{2}(1-x) + \frac{\beta}{2}(1+x)$. Em vez de se usar a base de Chebyshev $\{T_i\}_{i\geq 0}$ considera-se a base $\{\phi_i\}_{i\geq 2}$ definida por

$$\phi_i(x) = \begin{cases} T_i(x) - T_0(x), & \text{se } i \text{ par}, \\ \\ T_i(x) - T_1(x), & \text{se } i \text{ impar}. \end{cases} \tag{1.40}$$

Deste modo uma aproximação $v_N(x)$ de $v(x)$ toma a forma $v_N(x) := \sum_{i=2}^{N} v_i^{(N)} \phi_i(x)$. Para se determinar $v_N(x)$ escolhe-se para pontos de colocação os nós interiores de Chebyshev-Gauss-Lobato que para simplificar a notação representaremos por x_k, $1 \leq k \leq N-1$ onde (D.1),

$$x_k = \eta_k^{(N)} = -\cos\left(\frac{k\pi}{N}\right), \quad 1 \leq k \leq N-1.$$

Os nós extremos são desnecessários dado que os polinómios ϕ_i satisfazem as condições de fronteira de Dirichlet homogéneas.

Definindo os vetores

$$\mathbf{v} = \begin{bmatrix} v_2^{(N)} \\ v_3^{(N)} \\ \vdots \\ v_N^{(N)} \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g(x_1) \\ g(x_2) \\ \vdots \\ g(x_{N-1}) \end{bmatrix},$$

e a matriz $\mathbf{H} = (h_{ij})$, $1 \leq i, j \leq N-1$ com

$$h_{ij} = \left(\mathscr{L}\phi_{j+1}(x)\right)\big|_{x=x_i}, \quad 1 \leq i, j \leq N-1,$$

se a matriz $\mathbf{H}$ for regular então, o problema de encontrar a solução de colocação do problema (1.39) reduz-se à resolução de um sistema de $N-1$ equações lineares a $N-1$ incógnitas com forma matricial

$$\mathbf{Hv} = \mathbf{g}. \tag{1.41}$$

Finalmente, tendo em conta que $B(x) = \frac{\alpha+\beta}{2}T_0(x) + \frac{\beta-\alpha}{2}T_1(x)$, encontra-se a solução de colocação do problema (1.38) na base de Chebyshev, $u_N(x) = \sum_{i=0}^{N} u_i^{(N)} T_i(x)$ onde

$$u_0^{(N)} = \frac{\alpha+\beta}{2} + \sum_{i \text{ par}} v_i^{(N)}, \quad u_1^{(N)} = \frac{\beta-\alpha}{2} + \sum_{i \text{ impar}} v_i^{(N)}, \quad u_i^{(N)} = v_i^{(N)}, \quad i = 2, 3, \ldots, N.$$

Exemplo 1.4.2. Pretende-se encontrar aproximações, usando um método de colocação na base de Chebyshev, de funções u a duas variáveis, definidas pela equação de Poisson com condições fronteira de Dirichlet no quadrado $\Omega =]-1, 1[^2$. Mais exatamente tem-se que a função u é solução do problema

$$\mathscr{L}u = g, \quad \text{em } \Omega, \tag{1.42a}$$

$$\begin{aligned} u(-1,y) = h_W(y), \quad u(1,y) = h_E(y), \\ u(x,-1) = h_S(x), \quad u(x,1) = h_N(x), \end{aligned} \tag{1.42b}$$

onde

$$\mathscr{L}u := \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2},$$

e a função que determina a condição fronteira é suficientemente regular, i.e., as funções $h_W(y)$, $h_E(y)$, $h_S(x)$ e $h_N(x)$ pertencem ao domínio do operador linear $\mathscr{L}$ e definem uma função $h(x,y)$ contínua em $\overline{\Omega}$.

Analogamente ao exemplo 1.4.1 resolve-se o seguinte problema equivalente com condições fronteira homogéneas

$$\mathscr{L}v = g, \quad \text{em } \Omega, \tag{1.43a}$$

$$v(x,y) = 0, \quad \text{em } \overline{\Omega}, \tag{1.43b}$$

onde $v(x,y) = u(x,y) - B(x,y)$, $g(x,y) = f(x,y) - \mathscr{L}B(x,y)$ e a função B é determinada por *interpolação polinomial transfinita* [Boy01]

$$\begin{aligned} B(x,y) = &\frac{1-x}{2}h_W(y) + \frac{x+1}{2}h_E(y) + \frac{1-y}{2}h_S(x) + \frac{y+1}{2}h_N(x) \\ &- \frac{(1-x)(1-y)}{4}u(-1,-1) - \frac{(1-x)(y+1)}{4}u(-1,1) \\ &- \frac{(x+1)(1-y)}{4}u(1,-1) - \frac{(x+1)(y+1)}{4}u(1,1). \end{aligned}$$

Seja $v_N(x,y) = \sum_{i=2}^{N}\sum_{j=2}^{N} v_{ij}^{(N)}\Phi_{ij}(x,y)$ uma aproximação de colocação de uma solução do problema (1.43), com

$$\Phi_{ij}(x,y) \equiv \phi_i(x)\phi_j(y), \qquad 2 \leq i,j \leq N, \tag{1.44}$$

e as funções ϕ_k, $2 \leq k \leq N$ são os polinómios definidos em (1.40). Escolhe-se o conjunto C_N formado pelos nós de Chebyshev-Gauss-Lobatto que pertencem a Ω, i.e., considerando o conjunto de índices $J_e = \{(i,j)|\ 1 \leq i,j \leq N-1\}$ (aqui e no exemplo anterior tem-se $J = J_e$ e $J_b = \emptyset$) tem-se

$$C_N \equiv \left\{(x_i,y_j)|\ x_i = \eta_i^{(N)} \ \wedge \ y_j = \eta_j^{(N)}, (i,j) \in J\right\}.$$

A solução de colocação v_N anula o resíduo nos pontos de colocação, ou seja, verifica-se

$$\left(\mathscr{L}v_N - g\right)(x,y) = 0, \quad \text{em } C_N. \tag{1.45}$$

Ordenando o conjunto de índices J e, dado que as condições (1.45) fornecem $(N-1)^2$ equações e a função v_N possui $(N-1)^2$ coeficientes, $v_{ij}^{(N)}$, tem-se que a forma (1.37) reduz-se à resolução de um sistema de $(N-1)^2$ equações a $(N-1)^2$ incógnitas.

Contrariamente ao exemplo 1.4.1 a determinação da solução de colocação do problema (1.43) na base de Chebyshev $\{T_i(x)T_j(y)\}$, $0 \leq i, j \leq N$ não é imediata. Este facto deve-se a que, geralmente, são desconhecidos os desenvolvimentos na base de Chebyshev das funções $h_W(y)$, $h_E(y)$, $h_S(x)$ e $h_N(x)$ que definem as condições fronteira do problema. Contudo, observando que cada uma das quatro funções é solução de uma equação diferencial linear de segunda ordem com condições fronteira de Dirichlet, é possível encontrar uma aproximação para os coeficientes de Chebyshev destas funções. Por exemplo a função $h_W(y)$ é solução da equação

$$\frac{\mathrm{d}^2 u}{\mathrm{d}y^2} = g$$

com condições fronteira $u(-1) = h_W(-1)$ e $u(1) = h_W(1)$. Resolvendo esta equação, p. ex. usando o método de colocação descrito no exemplo 1.4.1, encontramos os coeficientes de Chebyshev aproximados de $h_W(y)$ e a passagem da parcela $\frac{1-x}{2}h_W(y)$ para a base é imediata. Procedendo deste modo para as restantes três funções obtém-se os coeficientes $b_{i,j}^{(N)}$ aproximados, da função

$$B(x,y) \approx \sum_{i=0}^{N}\sum_{j=0}^{N} b_{i,j}^{(N)} T_i(x)T_j(y).$$

Os testes efectuados, ver capítulo 4, mostram é suficiente calcular soluções de colocação, das quatro equações diferenciais ordinárias, de ordem N para obtermos uma boa aproximação dos coeficientes de Chebyshev de B.

Finalmente, para se determinar a solução v_N na base de Chebyshev, recorremos ao seguinte resultado,

Proposição 1.4.1. Seja $p(x,y) = \sum_{i=2}^{N}\sum_{j=2}^{N} a_{i,j}\Phi_{i,j}(x,y)$, onde $\Phi_{i,j}$ são os polinómios definidos em (1.44) então,

$$p(x,y) = \sum_{i=0}^{N}\sum_{j=0}^{N} b_{i,j}\mathcal{T}_{i,j}(x,y), \quad \text{com } \mathcal{T}_{i,j}(x,y) \equiv T_i(x)T_j(y)$$

onde,

$$
\begin{aligned}
b_{0,0} &= \sum_{i \text{ par}} \sum_{j \text{ par}} a_{i,j}, & b_{1,0} &= \sum_{i \text{ impar}} \sum_{j \text{ par}} a_{i,j}, \\
b_{0,1} &= \sum_{i \text{ par}} \sum_{j \text{ impar}} a_{i,j}, & b_{1,1} &= \sum_{i \text{ impar}} \sum_{j \text{ impar}} a_{i,j}, \\
b_{0,j} &= -\sum_{i \text{ par}} a_{i,j}, & b_{1,j} &= -\sum_{i \text{ impar}} a_{i,j}, \quad 2 \le j \le N, \\
b_{i,0} &= -\sum_{j \text{ par}} a_{i,j}, & b_{i,1} &= -\sum_{j \text{ impar}} a_{i,j}, \quad 2 \le i \le N, \\
b_{i,j} &= a_{i,j}, \quad \text{se } i \ge 2 \ \wedge \ j \ge 2.
\end{aligned}
\tag{1.46}
$$

Demonstração: *Como para $i, j \ge 2$ se tem,*

$$
\Phi_{i,j}(x,y) = \begin{cases} (\mathcal{T}_{i,j} - \mathcal{T}_{i,0} - \mathcal{T}_{0,j} + \mathcal{T}_{0,0}), & \textit{se } i \textit{ par e } j \textit{ par} \\ (\mathcal{T}_{i,j} - \mathcal{T}_{i,1} - \mathcal{T}_{0,j} + \mathcal{T}_{0,1}), & \textit{se } i \textit{ par e } j \textit{ impar} \\ (\mathcal{T}_{i,j} - \mathcal{T}_{i,0} - \mathcal{T}_{1,j} + \mathcal{T}_{1,0}), & \textit{se } i \textit{ impar e } j \textit{ par} \\ (\mathcal{T}_{i,j} - \mathcal{T}_{i,1} - \mathcal{T}_{1,j} + \mathcal{T}_{1,1}), & \textit{se } i \textit{ impar e } j \textit{ impar} \end{cases},
$$

somando e agrupando os índices correspondentes tem-se ,

$$
\begin{aligned}
\sum_{i=2}^{N}\sum_{j=2}^{N} a_{i,j}\Phi_{i,j} = &\left(\sum_{ipar}\sum_{j\ par} a_{i,j}\right)\mathcal{T}_{0,0} + \left(\sum_{i\ par}\sum_{j\ impar} a_{i,j}\right)\mathcal{T}_{0,1} + \\
&\left(\sum_{i\ impar}\sum_{j\ par} a_{i,j}\right)\mathcal{T}_{1,0} + \left(\sum_{i\ impar}\sum_{j\ impar} a_{i,j}\right)\mathcal{T}_{1,1} + \\
&\sum_{i=2}^{N}\left(-\sum_{j\ par} a_{i,j}\right)\mathcal{T}_{i,0} + \sum_{i=2}^{N}\left(-\sum_{j\ impar} a_{i,j}\right)\mathcal{T}_{i,1} + \\
&\sum_{j=2}^{N}\left(-\sum_{i\ par} a_{i,j}\right)\mathcal{T}_{0,j} + \sum_{j=2}^{N}\left(-\sum_{i\ impar} a_{i,j}\right)\mathcal{T}_{1,j} + \\
&\sum_{i=2}^{N}\sum_{j=2}^{N} a_{i,j}\mathcal{T}_{i,j}.
\end{aligned}
$$

❑

Supondo a solução de colocação do problema (1.43) escrita na base de Chebyshev $v_N(x,y) = \sum_{i=2}^{N}\sum_{j=2}^{N} v_{ij}^{(N)} T_i(x)T_j(y)$ pode-se escrever a solução de colocação, u_N do problema (1.42) na forma

$$
u_N(x,y) = \sum_{i=2}^{N}\sum_{j=2}^{N} v_{ij}^{(N)} T_i(x)T_j(y) + \sum_{i=0}^{N}\sum_{j=0}^{N} b_{i,j}^{(N)} T_i(x)T_j(y). \tag{1.47}
$$

Estabilidade e Convergência

Relativamente à estabilidade, iremos assumir que o operador linear $\mathscr{L}$ satisfaz as condições de coercividade e de continuidade num espaço E que satisfaz $X_N \subset E$ para

todo inteiro positivo N. Além disso existe $C > 0$ tal que para todo $N > 0$ e para todo $u \in X_N$ tem-se $\|u\|_N \leq C \|u\|_E$. Sob estas condições tem-se o seguinte

Teorema 1.4.2. [CHQZ07] Se existe uma constante $\bar{\alpha}$ para todo $N > 0$ tal que

$$\bar{\alpha} \|u\|_E^2 \leq (Q_N \mathscr{L}_N u, u), \quad \text{para todo } u \in X_N,$$

então a aproximação de colocação (1.36) é estável no sentido de que

$$\|u\|_E \leq \frac{C}{\bar{\alpha}} \|f\|_N .$$

1.5 Método Tau

Começamos com o caso em que o problema diferencial (1.1)-(1.2) está definido no domínio $\Omega =]-1, 1[$.

Seja $\{\phi_k\}_{k \in \mathbb{N}_0}$ um sistema de polinómios ortogonais relativamente a um produto interno $(u, v)_w = \int_{-1}^{1} u(x)v(x)w(x)dx$, onde $w > 0$ em Ω é a função peso e ϕ_k é um polinómio de grau k. A solução Tau de grau N é um polinómio $u_N = \sum_{k=0}^{N} c_k^{(N)} \phi_k$, onde os coeficientes $c_k^{(N)}$ são as incógnitas do problema. Seja β o número de condições fronteira, p.ex. $\beta = 2$ se $\mathscr{L}$ for um operador não degenerado de ordem 2, projeta-se a equação (1.1) no espaço dos polinómios de grau $N - \beta$,

$$(\mathscr{L} u_N, \phi_k)_w = (f, \phi_k)_w \tag{1.48}$$

e as condições fronteira (1.2) são impostas em $\partial\Omega_b$

$$\sum_{k=0}^{N} c_k^{(N)} \mathscr{B} \phi_k = 0, \quad \text{nos pontos de } \partial\Omega_b.$$

Seguindo o enquadramento da secção 1.1, fazemos $\mathbf{X} = L_w^2(]-1, 1[)$,

$$X_N = \{v \in \mathcal{P}_N \mid \mathscr{B} v = 0 \text{ nos pontos de } \partial\Omega_b\},$$

e

$$Y_N = \mathcal{P}_{N-\beta}.$$

Então o método Tau é equivalente a

$$\begin{aligned} &u_N \in X_N, \\ &(\mathscr{L} u_N, v) = (f, v) \quad \forall v \in Y_N. \end{aligned} \tag{1.49}$$

Relativamente à forma operacional (1.21), no método Tau o operador projecção Q_N projecta elementos de X no espaço Y_N relativamente ao produto interno (u, v) em X.

Para o caso multidimensional, consideramos que o problema diferencial tem domínio $\Omega = \prod_{k=1}^{d}]-1,1[$, $d \geq 1$, e que os polinómios $\mathcal{P}_N(\Omega)$ são algébricos em cada uma das d-variáveis, $x_1, \ldots, x_d$. Assumimos que em cada um dos lados L_i^c de Ω, onde $i \in \{1, \ldots, d\}$ e $c \in \{-1, 1\}$ sendo que

$$L_i^c = \{\mathbf{x} \equiv (x_1, \ldots x_d) \in \partial\Omega \mid x_i = c\},$$

apenas se tem um tipo de condições suplementares. Pode-se encontrar uma base de $\mathcal{P}_N(\Omega)$ usando o produto de funções base $\{\phi_k\}$ em cada variável. Definindo o reticulado

$$J = \{\mathbf{k} = (k_1, \ldots, k_n) \mid 0 \leq k_i \leq N \text{ para } i = 1, \ldots, d\},$$

e fazendo

$$\phi_{\mathbf{k}}(\mathbf{x}) = \phi_{k_1}(x_1) \cdots \phi_{k_d}(x_d).$$

Então $\{\phi_{\mathbf{k}},\ \mathbf{k} \in J\}$ é uma base para $\mathcal{P}_N(\Omega)$, com o produto interno

$$(u, v) = \int_{\Omega} u(\mathbf{x}) v(\mathbf{x}) w(\mathbf{x}) d\mathbf{x}$$

onde, $w(\mathbf{x}) = \prod_{i=1}^{d} w(x_i)$.

A solução Tau é um polinómio em $\mathcal{P}_N(\Omega)$ e os seus coeficientes são determinados resolvendo dois conjuntos de equações lineares. Seja β_i o número de condições suplementares estabelecidas nos lados $x_i = \pm 1$. Definindo o sub reticulado

$$J_e = \{\mathbf{k} = (k_1, \ldots, k_d) \in J \mid 0 \leq k_i \leq N - \beta_i, \quad \text{para } i = 1, \ldots, d\},$$

obtém-se o primeiro conjunto de equações lineares exigindo que a solução Tau $u_N \in \mathcal{P}_N(\Omega)$ satisfaça

$$(\mathscr{L} u_N, \phi_{\mathbf{k}}) = (f, \phi_{\mathbf{k}}), \quad \forall \mathbf{k} \in J_e. \tag{1.50}$$

O segundo conjunto obtém-se impondo as condições suplementares. Juntando todas as equações obtém-se um sistema de equações lineares algébricas cujas incógnitas são os coeficientes de u_N relativamente à base ortogonal $\{\phi_{\mathbf{k}} \mid \mathbf{k} \in J\}$.

Nas duas subsecções seguintes iremos descrever o algoritmo usado ao longo deste trabalho para encontrar aproximações Tau de soluções de equações diferenciais. Começamos por descrever a abordagem operacional do método Tau sugerida por Ortiz e Samara em [OS80] e [OS81], e na subsecção 1.5.2 apresentamos uma versão alterada, [MRMC], da abordagem operacional "clássica". Esta alteração foi efetuada de modo a obter um algoritmo numericamente mais estável.

1.5.1 Abordagem operacional clássica

A formulação operacional do método Tau baseia-se na representação matricial[1] dos operadores diferenciais na classe dos operadores diferenciais lineares de ordem finita, m, com coeficientes polinomiais a qual representaremos por $\mathfrak{L}$. Esta representação é igualmente extensível a operadores integrais, integro-diferenciais e a operadores vetoriais, [MRV04].

Representação Matricial de Operadores diferenciais em $\mathfrak{L}$:

Seja $\mathscr{L} \in \mathfrak{L}$ um operador definido por

$$\mathscr{L} \equiv \sum_{i=0}^{m} p_i(x) \frac{d^i}{dx^i}, \quad p_i(x) = \sum_{j=0}^{n_i} p_{i,j} x^j \equiv \mathbf{p_i} \cdot \mathbf{x} \in \mathcal{P}_{n_i}, \tag{1.51}$$

onde os polinómios $p_i(x)$, $i = 0, 1, \ldots, m$ são representados na forma matricial por produtos das matrizes infinitas, $\mathbf{x} = [1, x, x^2, \ldots]^T$, $\mathbf{p_i} = [p_{i,0}, p_{i,1}, \ldots, p_{i,n_i}, 0, 0, \ldots]$.

Dado um polinómio $y_n \in \mathcal{P}_n$, $y_n(x) = \sum_{k=0}^{n} a_k x^k$, representado na forma vetorial, por $\mathbf{y} = \mathbf{a} \cdot \mathbf{x}$, onde $\mathbf{a} = [a_0, a_1, \ldots, a_n, 0, 0, \ldots]$. Então,

$$\begin{aligned} \frac{d^k}{dx^k} y_n(x) &= \mathbf{a} \cdot \boldsymbol{\eta}^k \cdot \mathbf{x} \\ & \qquad\qquad , k = 0, 1, 2, \ldots \\ x^k y_n(x) &= \mathbf{a} \cdot \boldsymbol{\mu}^k \cdot \mathbf{x} \end{aligned}$$

onde,

$$\boldsymbol{\eta} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & 2 & 0 & \\ & & \cdots & \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & 0 & 1 \\ & & & \cdots \end{bmatrix}.$$

Definindo a matriz $\mathbf{\Pi}$,

$$\mathbf{\Pi} = \sum_{i=0}^{m} \boldsymbol{\eta}^i p_i(\boldsymbol{\mu}),$$

podemos representar a imagem de $y(x)$ pelo operador $\mathscr{L}$ na base das potências da seguinte forma [OS81],

$$\mathscr{L}\mathbf{y}(x) = \mathbf{a} \cdot \mathbf{\Pi} \cdot \mathbf{x}.$$

A matriz $\mathbf{\Pi}$ é a matriz infinita associada ao operador $\mathscr{L}$ e é necessariamente uma matriz banda cuja i-ésima linha é um vetor infinito que representa o *polinómio gerador* de ordem i do operador $\mathscr{L}$, $p_{\mathscr{L}}$, definido por $pol_i(x) = \mathscr{L}x^i$, para $i = 0, 1, 2, \ldots$. Definindo a *profundidade* do operador $\mathscr{L}$ como sendo o número $d = \min(b_i - i)$, onde b_i é o menor

[1] Iremos usar letras a negrito para a representação matricial de polinómios e operadores.

índice j tal que $p_{i,j} \neq 0$, e definindo a *altura* do operador $\mathscr{L}$ como sendo o número $h = \max_{n\in\mathbb{N}_0}\{m_n - n\}$, onde m_n é o grau do polinómio $\mathscr{L}x^n$, então as entradas não nulas da matriz $\mathbf{\Pi} = (\pi_{i,j})_{i,j\geq 0}$ estão todas na banda compreendida entre as paradiagonais $\pi_{i-d,i}$ e $\pi_{i,i+h-d+1}$.

Consideremos a matriz $\mathbf{v} \equiv [v_0(x), v_1(x), v_2(x), \ldots]^T$, onde $v_i(x)$, $i \geq 0$, são polinómios que verificam a condição, para cada $k \in \mathbb{N}_0$ é $\{v_i(x)|\ i = 0, 1, 2, \ldots k\}$ uma base de $\mathcal{P}_k$. Note que esta condição implica que, para todo o $\ell \in \mathbb{N}_0$ o polinómio $v_\ell(x)$ tem grau ℓ. Consideremos ainda a matriz mudança de base $\mathbf{V}$ definida pela condição

$$\mathbf{v} = \mathbf{V} \cdot \mathbf{x}.$$

Definindo a matriz

$$\mathbf{\Pi_v} = \mathbf{V} \cdot \mathbf{\Pi} \cdot \mathbf{V}^{-1},$$

então a imagem do polinómio $y(x)$ pelo operador $\mathscr{L}$ na base $\{v_k\}_{k\geq 0}$, é representada na forma matricial por

$$\mathscr{L}\mathbf{y} = \mathbf{a} \cdot \mathbf{\Pi_v} \cdot \mathbf{v}.$$

Formulação operacional do método Tau para operadores diferenciais em $\mathfrak{L}$:

Consideremos a seguinte equação diferencial

$$\begin{array}{l} \mathscr{L}u = f, \quad \text{em } \Omega =]a, b[. \\ g_j(u) = \sigma_j, \quad j = 1, 2, \ldots, m \end{array} \tag{1.52}$$

onde $\mathscr{L} \in \mathfrak{L}$, g_j, $j = 1, 2, \ldots, m$ são funcionais lineares, que representam as condições suplementares e $f(x) = \sum_{i=0}^{\ell} f_i x^i$ é um polinómio de grau ℓ.

A aproximação Tau u_N de grau N, escrita numa base $\{v_k\}_{k\geq 0}$, é a solução do problema Tau associado ao problema (1.52)

$$\begin{array}{l} \mathscr{L}u(x) = f(x) + H_N(x), \quad x \in \Omega \\ g_j(u) = \sigma_j, \quad j = 1, 2, \ldots, m, \end{array} \tag{1.53}$$

onde H_N é um certo polinómio chamado de *perturbação* ou *ruído* do problema Tau.

Em termos matriciais, considerando:

$$u_N(x) = \boldsymbol{\alpha}^{(N)} \cdot \mathbf{v}, \quad \boldsymbol{\alpha}^{(N)} = \left[\alpha_0^{(N)}, \alpha_1^{(N)}, \cdots, \alpha_N^{(N)}, 0, 0, \cdots\right],$$

a matriz

$$\mathbf{B} = (\beta_{i,j}), \quad \beta_{i,j} = \begin{cases} g_j(x^i),\ j = 1, 2, \ldots, m \ \text{ e } \ i = 0, 1, 2, \ldots \\ 0, \quad j > m \end{cases},$$

e os vetores

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_m, 0, 0, \ldots) \quad \text{e} \quad \mathbf{f} = (f_1, f_2, \ldots, f_\ell, 0, 0, \ldots)$$

e definindo as matrizes

$$\begin{aligned} \mathbf{B_v} &= \mathbf{V} \cdot \mathbf{B} \\ \mathbf{f_v} &= \mathbf{f} \cdot \mathbf{V}^{-1} \\ \boldsymbol{\Gamma_v} &= \mathbf{B_v} + \boldsymbol{\Pi_v} \cdot \boldsymbol{\mu}^m \\ \boldsymbol{\beta} &= \boldsymbol{\sigma} + \mathbf{f_v} \cdot \boldsymbol{\mu}^m \end{aligned}$$

então, sendo $u = \boldsymbol{\alpha} \cdot \mathbf{v}$ o desenvolvimento formal de u na base $\{v_k(x)\}_{k\geq 0}$, $\boldsymbol{\alpha}$ é solução do sistema infinito

$$\boldsymbol{\alpha} \cdot \boldsymbol{\Gamma}_v = \boldsymbol{\beta}$$

o qual, reduzido às suas primeiras $N+1$ equações,

$$\boldsymbol{\alpha}^{(\boldsymbol{N})} \cdot \boldsymbol{\Gamma}_{\mathbf{v}}^{(\mathbf{N})} = \boldsymbol{\beta}^{(\boldsymbol{N})}, \tag{1.54}$$

conduz à aproximação Tau [OS81]

$$u_N(x) = \boldsymbol{\alpha}^{(\boldsymbol{N})} \cdot \mathbf{v}^{(\mathbf{N})}$$

com *perturbação Tau*

$$H_N(x) = \sum_{i=1}^{m+h} \tau_i^{(N)} v_{N-m+i}(x), \quad \tau_i^{(N)} = \boldsymbol{\alpha}^{(\boldsymbol{N})} \boldsymbol{\Pi}_{\mathbf{v}}^{(\mathbf{N})} \cdot \mathbf{e}_{n-m+i},$$

onde $\mathbf{e}_j$ designa o vetor com entrada unitária na j-ésima linha e restantes entradas nulas.

Esta formulação possui a vantagem de poder usar diferentes famílias de polinómios ortogonais e de tratar problemas com diferentes condições suplementares. Contudo, a sua implementação, exige que as operações de desvio e de derivação, sobre polinómios, sejam efetuadas na base das potências. Estas mudanças de base implicam, na implementação matricial, o cálculo da inversa da matriz $\mathbf{V}$. Como, as matrizes $\mathbf{V}$ são, para dimensões elevadas, mal condicionadas os algoritmos ficam numericamente instáveis, [MRMC]. Este problema é especialmente relevante quando o problema a resolver tem solução cujas soluções Tau têm convergência lenta. Estas observações motivaram as seguintes alterações à abordagem operacional sugeridas em [MRMC].

1.5.2 Abordagem operacional modificada

Suponhamos que a base $\{v_k\}_{k\geq 0}$ do espaço $\mathcal{P}$, referida atrás, é constituída por um sistema de polinómios ortogonais num intervalo I relativamente a uma função peso w. Consideramos em $\mathcal{P}$ o produto interno usual $(p,q)_w = \int_I pqwdx$, para todos $p, q \in \mathcal{P}$ e norma associada $\|p\|_w \equiv (p,q)_w^{1/2}$.

A representação, na forma matricial, dos operadores de desvio e de derivação para polinómios na base $\{v_k\}_{k\geq 0}$ pode ser efetuada usando as matrizes $\boldsymbol{\mu}_{\mathbf{v}} \equiv [\mu_{i,j}]_{i,j\geq 0}$ e $\boldsymbol{\eta}_{\mathbf{v}} \equiv [\eta_{i,j}]_{i,j\geq 0}$ definidas, respetivamente, pelas condições

$$\mathbf{x}\cdot\mathbf{v} = \boldsymbol{\mu}_{\mathbf{v}}\mathbf{v}, \quad \text{e} \quad \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{v} = \boldsymbol{\eta}_{\mathbf{v}}\mathbf{v}.$$

Com estas definições, temos o seguinte

Teorema 1.5.1. Dado um polinómio $y(x) = \sum_{k=0}^{n} a_k v_k$ com representação matricial $\mathbf{y} = \mathbf{a}_{\mathbf{v}}\cdot\mathbf{v}$ e o operador $\mathscr{L}$ definido por (1.51) tem-se as seguintes representações matriciais

1. $x^k y(x) = \mathbf{a}_{\mathbf{v}} \cdot {\boldsymbol{\mu}_{\mathbf{v}}}^k \cdot \mathbf{v}, \quad k = 0, 1, \ldots,$
2. $\dfrac{\mathrm{d}^k}{\mathrm{d}x} y(x) = \mathbf{a}_{\mathbf{v}} \cdot {\boldsymbol{\eta}_{\mathbf{v}}}^k \cdot \mathbf{v}, \quad k = 0, 1, \ldots,$
3. $\mathscr{L} y(x) = \mathbf{a}_{\mathbf{v}} \cdot \boldsymbol{\Pi}_{\mathbf{v}} \cdot \mathbf{v}, \quad \text{onde,} \quad \boldsymbol{\Pi}_{\mathbf{v}} = \displaystyle\sum_{k=0}^{m} \boldsymbol{\eta}_{\mathbf{v}}^k p_k\left(\boldsymbol{\mu}_{\mathbf{v}}\right).$

As entradas das matrizes $\boldsymbol{\mu}_{\mathbf{v}}$ e $\boldsymbol{\eta}_{\mathbf{v}}$ podem ser calculadas, respetivamente, pelas relações

$$\mu_{i,j} = \frac{(xv_i, v_j)_w}{\|v_i\|_w} \quad \text{e} \quad \eta_{i,j} = \frac{(v_i', v_j)_w}{\|v_i\|_w} \quad i,j \geq 0. \tag{1.55}$$

Contudo, as igualdades (1.55) não são o modo mais eficiente para se determinar as matrizes $\boldsymbol{\mu}_{\mathbf{v}}$ e $\boldsymbol{\eta}_{\mathbf{v}}$. Os polinómios $\{v_k\}_{k\geq 0}$ formam um sistema ortogonal então pode-se colocar a relação de recorrência (A.2) na forma

$$\begin{cases} xv_k = \alpha_k v_{k+1} + \beta_k v_k + \gamma_k v_{k-1}, \quad k = 0, 1, 2 \ldots \\ v_{-1} = 0, \quad v_0 = 1 \end{cases} \tag{1.56}$$

e as entradas das matrizes $\boldsymbol{\mu}_{\mathbf{v}}$ e $\boldsymbol{\eta}_{\mathbf{v}}$ podem ser calculadas usando o seguinte

Teorema 1.5.2. Seja $\{v_k\}_{k\geq 0}$ uma sucessão de polinómios ortogonais determinados pela relação de recorrência (1.56) então,

$$\boldsymbol{\mu}_{\mathbf{v}} = \begin{bmatrix} \beta_0 & \alpha_0 & & \\ \gamma_1 & \beta_1 & \alpha_1 & \\ & \gamma_2 & \beta_2 & \alpha_2 \\ & & & \cdots \end{bmatrix} \tag{1.57}$$

$$\boldsymbol{\eta}_{\mathbf{v}} = \begin{bmatrix} 0 & & & \\ \eta_{1,0} & 0 & & \\ \eta_{2,0} & \eta_{2,1} & 0 & \\ \eta_{3,0} & \eta_{3,1} & \eta_{3,2} & 0 \\ & & & \cdots \end{bmatrix}, \tag{1.58}$$

onde para cada $i \geq 1$

$$\begin{cases} \eta_{i+1,j} = \dfrac{1}{\alpha_i}\left[\alpha_{j-1}\eta_{i,j-1} + (\beta_j - \beta_i)\,\eta_{i,j} + \gamma_{i+1}\eta_{i,j+1} - \gamma_i\eta_{i-1,j}\right], \ j = 0, 1, \ldots, i-1 \\ \eta_{i+1,i} = \dfrac{1}{\alpha_i}\left(\alpha_{i-1}\eta_{i,i-1} + 1\right). \end{cases}$$

Iremos, de seguida, particularizar o cálculo das matrizes $\boldsymbol{\mu}_{\mathbf{v}}$ e $\boldsymbol{\eta}_{\mathbf{v}}$ para os polinómios de Chebyshev e de Legendre.

Polinómios de Chebyshev:

Para os polinómios de Chebyshev $\{T_k\}_{k\geq 0}$ a relação (1.56) toma a forma

$$\begin{cases} xT_k = \frac{1}{2}T_{k+1} + \frac{1}{2}T_{k-1}, \quad k = 0, 1, 2\ldots \\ T_{-1} = 0, \quad T_0 = 1 \end{cases} \tag{1.59}$$

e como [AS65]

$$\frac{\mathrm{d}T_{2k}(x)}{\mathrm{d}x} = 4k\sum_{i=1}^{k} T_{2i-1}(x), \quad \text{e,} \quad \frac{\mathrm{d}T_{2k+1}(x)}{\mathrm{d}x} = (2k+1) + 2(2k+1)\sum_{i=1}^{k} T_{2i}(x)$$

tem-se

$$\boldsymbol{\mu}_{\mathbf{T}} = \begin{bmatrix} 0 & 1 & & & \\ 1/2 & 0 & 1/2 & & \\ & 1/2 & 0 & 1/2 & \\ & & & \cdots & \end{bmatrix} \quad \text{e} \quad \boldsymbol{\eta}_{\mathbf{T}} = \begin{bmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ 0 & 4 & 0 & & & \\ 3 & 0 & 6 & 0 & & \\ 0 & 8 & 0 & 8 & 0 & \\ 5 & 0 & 10 & 0 & 10 & 0 \\ & & & & & \cdots \end{bmatrix},$$

ou seja, tem-se $\eta_{i,j} = 0$ exceto nos seguintes casos

$$\begin{cases} \eta_{i,j} = 2i, \quad j = i-1 : -2 : 1, \ i \geq 1 \\ \eta_{2i+1,0} = 2i+1, \ i > 0 \end{cases}$$

com a convenção de que $j = n : -k : m$, com $k > 0$ e $n > m$, significa que j toma os valores $j = n - k\ell$, $\ell = 0, 1, \ldots, k_m$, onde k_m é o maior valor tal que $j < m$.

Polinómios de Legendre:

Para os polinómios de Legendre $\{P_k\}_{k\geq 0}$ a relação (1.56) toma a forma

$$\begin{cases} xP_k = \dfrac{k+2}{2k+1}P_{k+1} + \dfrac{k}{2k+1}P_{k-1}, \quad k = 0, 1, 2\ldots \\ P_{-1} = 0, \quad P_0 = 1 \end{cases} \tag{1.60}$$

e

$$\frac{\mathrm{d}P_{2k}(x)}{\mathrm{d}x} = \sum_{i=1}^{k}(4i-1)P_{2i-1}(x), \quad \text{e,} \quad \frac{\mathrm{d}P_{2k+1}(x)}{\mathrm{d}x} = \sum_{i=0}^{k}(4i+1)P_{2i}(x).$$

Então, tem-se $\eta_{i,j} = 0$ exceto para

$$\eta_{i,j} = 2j+1, \quad j = i-1:-2:0, \ i \geq 1.$$

Consequentemente tem-se

$$\boldsymbol{\mu}_{\mathbf{P}} = \begin{bmatrix} 0 & 1 & & \\ 1/3 & 0 & 2/3 & \\ & 2/5 & 0 & 3/5 \\ & & & \cdots \end{bmatrix} \quad \text{e} \quad \boldsymbol{\eta}_{\mathbf{P}} = \begin{bmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ 0 & 3 & 0 & & & \\ 1 & 0 & 5 & 0 & & \\ 0 & 3 & 0 & 7 & 0 & \\ 1 & 0 & 5 & 0 & 9 & 0 \\ & & & & & \cdots \end{bmatrix},$$

Exemplo 1.5.1. Consideremos os operadores diferenciais lineares $\mathscr{L}_\alpha$, definidos por

$$\mathscr{L}_\alpha := \left(x - \frac{\alpha^2+1}{2\alpha}\right)\frac{d^2}{dx^2} + \frac{d}{dx}, \quad \alpha \in \mathbb{R} \setminus \{-1, 0, 1\}.$$

Pretendemos encontrar aproximações Tau $y_N^{(\alpha)}(x)$ da solução exata $y^{(\alpha)}(x)$ do problema diferencial

$$\mathscr{L}_\alpha y^{(\alpha)}(x) = 0, \quad -1 < x < 1 \tag{1.61}$$

sujeito às condições fronteira de Dirichlet $y^{(\alpha)}(-1) = 1 - \frac{1}{2}\log(1+2\alpha+\alpha^2)$, $y^{(\alpha)}(1) = 1 - \frac{1}{2}\log(1-2\alpha+\alpha^2)$. É conhecida a solução exata

$$y^{(\alpha)}(x) = 1 - \frac{1}{2}\log\big(1-2\alpha x+\alpha^2\big), \quad \text{onde} \quad \begin{cases} x < \frac{\alpha^2+1}{2\alpha}, & \alpha > 0 \\ \\ x > \frac{\alpha^2+1}{2\alpha}, & \alpha < 0 \end{cases} \tag{1.62}$$

e o desenvolvimento de Fourier na base dos polinómios de Chebyshev (de primeira espécie) ortogonais no intervalo $[-1, 1]$ [AS65]

$$y^{(\alpha)}(x) = \sum_{k=0}^{\infty} c_k(\alpha)T_k(x) = 1 + \sum_{k=1}^{\infty}\frac{\alpha^k}{k}T_k(x), \quad -1 < x < 1, \quad |\alpha| < 1.$$

Os polinómios geradores do operador $\mathscr{L}_\alpha$ são

$$\begin{aligned} pol_0(x) &= \mathscr{L}_\alpha 1 = 0 \\ pol_1(x) &= \mathscr{L}_\alpha x = 1 \\ pol_n(x) &= \mathscr{L}_\alpha x^n = -\tfrac{\alpha^2+1}{2\alpha}n(n-1)x^{n-2} + n^2x^{n-1}, \quad n \geq 2. \end{aligned}$$

então tem-se que o operador $\mathscr{L}_\alpha$ tem altura $h = \max\{0, -1\} = 0$ e é representado pela matriz

$$\mathbf{\Pi}_\alpha = \begin{bmatrix} 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & \cdots \\ -\frac{\alpha^2+1}{\alpha} & 4 & 0 & \cdots \\ & \ddots & \ddots & \\ & & -\frac{\alpha^2+1}{2\alpha} n(n-1) & n^2 & \cdots \\ & & & \cdots & \cdots \end{bmatrix}$$

e tem-se:

$$\mathbf{V} = \begin{bmatrix} 1 & & & & & & \\ 0 & 1 & & & & & \\ -1 & 0 & 2 & & & & \\ 0 & -3 & 0 & 4 & & & \\ 1 & 0 & -8 & 0 & 8 & & \\ 0 & 5 & 0 & -20 & 0 & 16 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 1 & 0 & \cdots \\ -1 & 1 & 0 & \cdots \\ 1 & 1 & 0 & \cdots \\ -1 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \\ (-1)^{n+1} & 1 & 0 & \cdots \\ & & \cdots & \cdots \end{bmatrix},$$

$$\boldsymbol{\sigma}_\alpha = \begin{bmatrix} 1 - 1/2\log(1 + 2\alpha + \alpha^2) \\ 1 - 1/2\log(1 - 2\alpha + \alpha^2) \\ 0 \\ \vdots \end{bmatrix}^T, \quad \text{e} \quad \mathbf{f} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}^T.$$

A função $y^{(\alpha)}$, $\alpha \neq 0$, é analítica em $\mathbb{C} \setminus]-\infty, \frac{\alpha^2+1}{2\alpha}[$ para $\alpha > 0$ e em $\mathbb{C} \setminus]\frac{\alpha^2+1}{2\alpha}, \infty[$ para $\alpha < 0$. Se $\alpha \in]-1, 1[\setminus\{0\}$, a série de Chebyshev da função $y^{(\alpha)}$ possui índice de convergência exponencial

$$r = \lim_{k\to\infty} \frac{\log|\log|c_k(\alpha)||}{\log k} = 1$$

com taxa de convergência assimptótica [Boy01]

$$\mu = \lim_{k\to\infty} -\frac{\log|c_k(\alpha)|}{k} = -\log|\alpha|.$$

Deste modo, a série de Chebyshev da função $y^{(\alpha)}$ converge mais rapidamente para valores de α próximos de zero e converge mais lentamente para valores de α próximos dos extremos do intervalo. Tomando para aproximação da função $y^{(\alpha)}$ a solução Tau de ordem N do problema (1.61)

$$y_N^{(\alpha)}(x) = \sum_{k=0}^{N} c_k^{(N)}(\alpha) T_k(x),$$

obteve-se um resultado semelhante. De facto, a convergência do método Tau é mais rápida para valores de α próximos de zero, como se pode observar pelos resultados indicados na Figura 1.1. Analogamente o erro nos coeficientes $\left|c_k(\alpha) - c_k^N(\alpha)\right|$ reflete os resultados verificados nos erros das soluções tau, o que era de esperar, dado que as séries de Chebyshev são convergentes, ver Figura 1.2. Outro facto verificado, é que os coeficientes das soluções tau obtidos, satisfazem as relações

$$c_k^N(-\alpha) = \begin{cases} c_k^N(\alpha), & k \text{ par} \\ -c_k^N(\alpha), & k \text{ impar} \end{cases}.$$

Logo, tendo em conta as simetrias existentes nos polinómios de Chebyshev, as aproximações $y_N^{(\alpha)}$ verificam igualmente a relação existente entre duas soluções do problema (1.61) para valores de α simétricos

$$y_N^{(-\alpha)}(x) = y_N^{(\alpha)}(-x).$$

1.5.3 Estabilidade e Convergência

Nesta secção iremos estudar a estabilidade e a convergência do método Tau. O caso mais simples ocorre quando o operador $\mathscr{L}$ é da forma $\mathscr{L} \equiv \frac{d^m}{dx^m} + \sum_{i=0}^{m-1} p_i(x)\frac{d^i}{dx^i}$, onde as funções coeficientes $p_i(x) \in L_w^2(I)$, $i = 0, \ldots, m-1$, $I =]a, b[$ e as condições suplementares são caracterizadas por m funcionais lineares B_i. Para funções coeficientes $p_i(x)$ polinomiais tem-se o seguinte resultado

Teorema 1.5.3. [RP89] Considere-se a equação diferencial

$$\begin{aligned} &\mathscr{L}u = 0, \quad \text{em } I, \\ &g_i u = \sigma_i, \quad i = 1, \ldots, m \end{aligned}$$

Se o problema homogéneo

$$\begin{aligned} &\mathscr{L}u = 0, \quad \text{em } I, \\ &g_i u = 0, \quad i = 1, \ldots, m \end{aligned}$$

tem apenas a solução nula, e as condições suplementares $g_i u = \sigma_i, \quad i = 1, \ldots, m$ satisfizerem a condição

$$\left\| \{g_i\}_{i=1,\ldots,m} \right\| \leq \text{const} \left\| u \right\|_{H_w^m(I)},$$

então a abordagem operacional do método tau conduz, para valores de N suficientemente grandes, a uma única solução que converge para a única solução do problema não homogéneo, relativamente à norma $\|*\|_{H_w^m(I)}$. Além disso os erros, na norma $\|*\|_{H_w^m(I)}$, cometidos pela solução tau u_N e a melhor aproximação da solução em $\mathcal{P}_N$ têm a mesma ordem de grandeza.

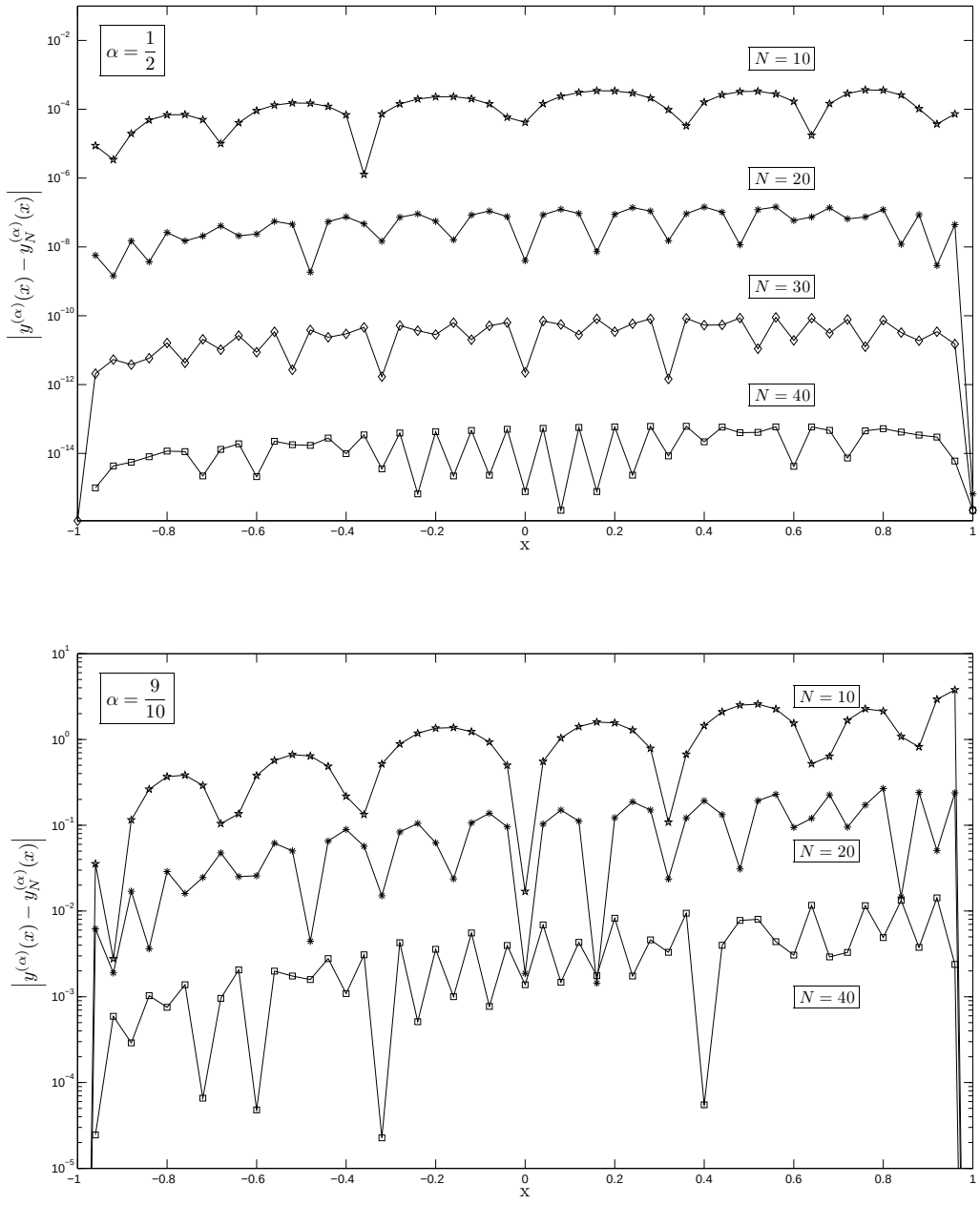


Figura 1.1: Erros absolutos de soluções Tau do problema (1.61) para valores de $\alpha = 1/2$ (em cima) e de $\alpha = 9/10$ (em baixo). O método tau converge claramente mais rapidamente para $\alpha = 1/2$.

Este resultado foi generalizado para operadores lineares com funções coeficientes $p_i(x) \in L_w^2(I)$ em [Cab94].

Contudo este resultado não é válido para problemas em derivadas parciais. Como o espaço X_N das funções base é diferente do espaço Y_N das funções teste, seguiremos a

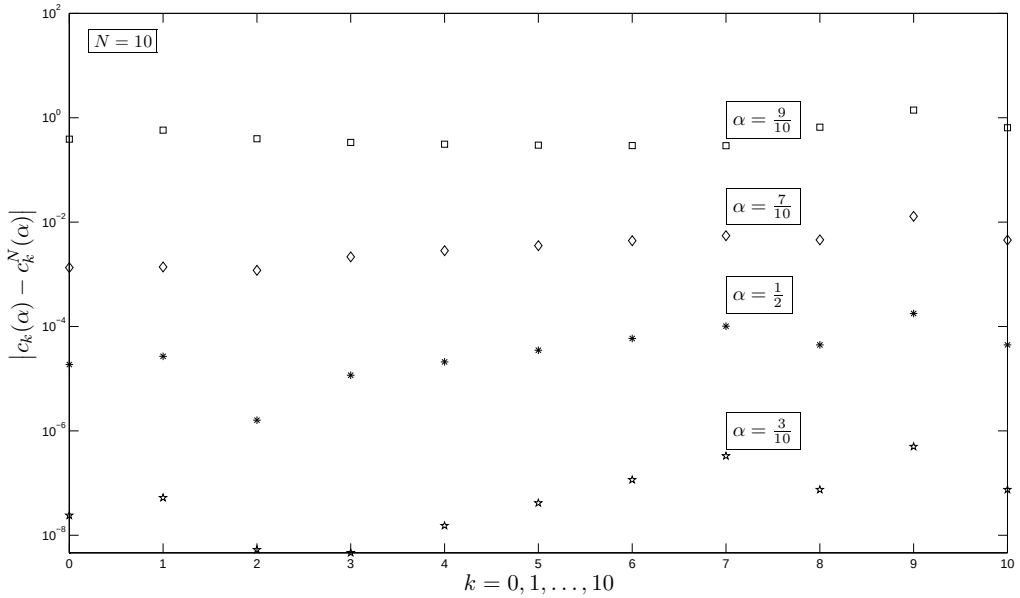


Figura 1.2: Erros absolutos dos coeficientes das soluções Tau de ordem 10 do problema (1.61) para diferentes valores de α.

abordagem da forma discreta da condição inf-sup (ver secção 1.1). Assumimos que o operador $\mathscr{L}$ está associado com uma forma bilinear $a(u, v) = (\mathscr{L}u, v)$ que satisfaz as condições (1.11)-(1.13). Adicionalmente assumiremos que para todo o $N \in \mathbb{N}_0$ tem-se $X_N \subset W$ e $Y_N \subset V$. Então, tem-se a seguinte condição inf-sup, devida a Babušca [BA72] que é a congénere discreta de (1.12).

Se existe uma constante $\bar{\alpha} > 0$ independente de N tal que

$$\bar{\alpha}||u||_W \leq \sup_{v \in Y_N \setminus \{0\}} \frac{(\mathscr{L}u, v)}{||v||_V}, \quad \forall u \in X_N, \tag{1.63}$$

então tem-se a seguinte estimativa

$$\|u_N\|_W \leq \frac{C}{\bar{\alpha}} \|f\|, \tag{1.64}$$

onde a constante C, independente de N, satisfaz a relação

$$||v|| \leq C||v||_V, \quad \forall v \in V. \tag{1.65}$$

Como os espaços X_N e Y_N têm a mesma dimensão a desigualdade (1.64) implica que o problema (1.49) tem solução única e que o método é estável. Dividindo ambos os membros de (1.49) por $||v||_V$ e tomando o supremo sobre todo o elemento v de Y_N e usando (1.63) mais a continuidade da inclusão de V em X obtêm-se o limite (1.64).

Relativamente à convergência do método. Seja R_N um operador linear de um espaço denso $\mathcal{W} \subset D_B(\mathscr{L})$ em X_N ($\mathcal{W}$ será um espaço de funções com certas condições de regularidade, p.e. das funções em $D_B(\mathscr{L})$ tais que são de classe $C^n(\Omega)$, para um certo $n \in \mathbb{N}_0$) tal que

$$\lim_{N\to\infty} \|u - R_N u\|_W = 0, \quad \forall u \in \mathcal{W}. \tag{1.66}$$

Sob esta hipótese de consistência a aproximação (1.49) é convergente e tem-se o seguinte limite para o erro da solução tau do problema (1.49) [CHQZ07]

$$\|u - u_N\|_W \leq \left(1 + \frac{A}{\bar{\alpha}}\right) \|u - R_N u\|_W \tag{1.67}$$

o que implica a convergência do método.

Exemplo 1.5.2. [CHQZ07] Consideremos o seguinte problema de valores fronteira de Dirichelet

$$\begin{aligned} &\mathscr{L}u \equiv -\tfrac{d^2u}{dx^2} + \lambda^2 u = f, \quad x \in I, \ \lambda \in \mathbb{R} \\ &u(-1) = u(1) = 0. \end{aligned} \tag{1.68}$$

onde $I =]-1, 1[$. Pretendemos determinar a solução tau expandida nos polinómios de Chebyshev. Assumimos que $f \in L^2_w(I)$, com $w(x) = (1-x^2)^{-1/2}$. Determinamos a solução $u_N(x) = \sum_{k=0}^{N} \hat{c}_k T_k(x)$ usando as equações

$$\int_{-1}^{1} \left[\left(-\frac{d^2u_N}{dx^2} + \lambda^2 u_N(x)\right) T_k(x) w(x)\right] dx = \int_{-1}^{1} f(x) T_k(x) w(x) dx, \ k = 0, 1, \ldots, N-2,$$

$$\sum_{k=0}^{N} (-1)^k \hat{c}_k = \sum_{k=0}^{N} \hat{c}_k = 0.$$

Neste exemplo temos, $X_N = \{v \in \mathcal{P}_N \mid v(-1) = v(1) = 0\}$ e $Y_N = \mathcal{P}_{N-2}$. Se $u \in X_N$ é um polinómio de grau N, então $v = -\frac{d^2u}{dx^2}$ é um polinómio de grau $N-2$ e tem-se

$$(\mathscr{L}u, v) = \int_{-1}^{1} \left(\frac{d^2u}{dx^2}\right)^2 w dx + \lambda^2 \int_{-1}^{1} \frac{du}{dx} \frac{d(uw)}{dx} dx.$$

Usando a desigualdade [CHQZ07]

$$\int_{-1}^{1} \frac{du}{dx} \frac{d(uw)}{dx} \geq \frac{1}{4} \left\| \frac{du}{dx} \right\|^2_{L^2_w(I)}, \quad \forall u \in H^1_{0,w}(I), \tag{1.69}$$

e a desigualdade de Poincaré (B.4.1) tem-se

$$(\mathscr{L}u, v) \geq \left\| \frac{d^2u}{dx^2} \right\|^2_{H^0_w(I)} + \frac{\lambda^2}{4} \|u\|^2_{H^0_w(I)} \geq C(I) \|u\|^2_{H^2_w(I)}.$$

Logo se escolhermos $W = H^2_{0,w}$ e $V = L^2_w(I)$ a condição inf-sup (1.63) é satisfeita, e tem-se

$$\|u_N\|_{H^2_w(I)} \leq C \|f\|_{H^0_w(I)} \tag{1.70}$$

para uma constante C independente de N e de λ.

A convergência pode estabelecer-se usando (1.67) e definindo o operador projecção R_N da seguinte forma. Seja u a solução exacta e seja $R_N u$ um polinómio algébrico de grau não superior a N que, para $0 \leq k \leq 2$, verifica

$$\|u - u_N\|_{H^k_w(I)} \leq C N^{k-m} |u|_{H^{m;N}_w(I)}$$

e que se anula nos extremos de I. A projecção R_N pode construir-se da forma $R_N u = P^2_N u - p_1$, onde $P^2_N u$ é a projecção ortogonal de u sobre $\mathcal{P}_N$ relativamente ao produto interno de $H^2_w(I)$, o qual satisfaz (1.67) e p_1 é um polinómio de grau 1 que toma os valores de $P^2_N u$ nos extremos de I. Tendo em conta a inclusão contínua $H^1_w(I) \subset C^0(I)$ [Ada78], tem-se

$$\|p_1\|_{H^2_w(I)} \leq C \left\|u - P^2_N u\right\|_{H^2_w(I)} \leq C N^{2-m} |u|_{H^{m;N}_w(I)}, \quad m \geq 2.$$

Obtém-se deste modo a estimativa de convergência

$$\|u - u_N\|_{H^2_w(I)} \leq C N^{2-m} |u|_{H^{m;N}_w(I)}, \quad m \geq 2. \tag{1.71}$$

Nota: *Podia-se usar para funções testes outros polinómios para obter a estabilidade deste esquema. Em vez de se utilizar* $v = -\frac{d^2 u_n}{dx^2}$ *podia-se usar* $v = P_{N-2}u$*, onde* u *é um polinómio qualquer em* X_N *e teriamos*

$$\begin{aligned} (\mathscr{L}u, v) &= -\int_{-1}^{1} \frac{d^2 u}{dx^2} P_{N-2} u w dx + \lambda^2 \int_{-1}^{1} u P_{N-2} u w dx \\ &= \int_{-1}^{1} \frac{du}{dx} \frac{d(uw)}{dx} dx + \lambda^2 \int_{-1}^{1} (P_{N-2}u)^2 w dx. \end{aligned}$$

Logo, o esquema tau verifica a estimativa

$$\frac{1}{2} \left\| \frac{du_N}{dx} \right\|_{H^0_w(I)} + \lambda \|P_{N-2} u_N\|_{H^0_w(I)} \leq C \|f\|_{H^0_w(I)}, \tag{1.72}$$

e se $\lambda \gg 1$*, tem-se que* $\|P_{N-2}u_N\|_{L^2_w(I)} \approx \mathcal{O}(1/\lambda)$.

Capítulo 2

Aproximação de Padé

Este capítulo é dedicado ao estudo da aproximação de Padé e divide-se essencialmente em duas partes. Na primeira parte, resumimos os resultados teóricos e alguns resultados sobre o cálculo numérico dos aproximantes de Padé de séries de potências. Na segunda parte, descrevemos a aproximação de Padé de séries de polinómios ortogonais.

2.1 Noções básicas sobre sucessões assimptóticas

Definição 2.1.1. Sejam $x_0 \geq 0$ e f, g duas funções definidas num intervalo I, limitado $I =]x_0, x_0 + \delta[$, $\delta > 0$, se x_0 for finito ou $I =]A, x_0[$, $A > 0$, se x_0 for infinito. Diz-se que:

(i) *a ordem da função f não excede a ordem da função g em x_0* e escreve-se $f(x) = \mathcal{O}(g(x))$ quando $x \to x_0$ se,

$$\left|\frac{f(x)}{g(x)}\right| \text{ é limitada no intervalo } I,$$

(ii) *a ordem da função f é inferior à ordem da função g em x_0* e escreve-se $f(x) = o(g(x))$ quando $x \to x_0$ se,

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0,$$

(iii) *a função f é assimptoticamente igual à função g em x_0* e escreve-se $f(x) \sim g(x)$ quando $x \to x_0$ se,

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 1,$$

onde, $\lim_{x \to x_0}$ é usado para representar os limites $\lim_{x \to x_0^+}$ ou $\lim_{x \to \infty}$ consoante x_0 seja finito ou infinito, respetivamente.

As definições 2.1.1 implicam algumas propriedades importantes, as quais resumimos na seguinte proposição.

Proposição 2.1.1. Dadas as definições 2.1.1 verificam-se as seguintes propriedades [Sid03]:

1. se $f = o(g)$ então $f = \mathcal{O}(g)$, contudo a conversa não se verifica.
2. $f = \mathcal{O}(g)$ não implica que $g = \mathcal{O}(f)$.
3. se $f = \mathcal{O}(g)$ e $g = \mathcal{O}(f)$ então $1/f = \mathcal{O}(1/g)$.
4. se $f \sim g$ então tem-se: $g \sim f$, $1/f \sim 1/g$, $f = \mathcal{O}(g)$, $g = \mathcal{O}(f)$, e (pela propriedade anterior) $1/f = \mathcal{O}(1/g)$ e $1/g = \mathcal{O}(1/f)$.

As definições 2.1.1 possuem versões análogas para sucessões.

Definição 2.1.2. Sejam $\{u_n\}$ e $\{v_n\}$ duas sucessões. Diz-se que:

(i) a ordem da sucessão $\{u_n\}$ não excede a ordem de $\{v_n\}$ e escreve-se $u_n = \mathcal{O}(v_n)$ se,

$$\left|\frac{u_n}{v_n}\right| \text{ é limitada,}$$

(ii) a ordem da sucessão $\{u_n\}$ é inferior à ordem de $\{v_n\}$ e escreve-se $u_n = o(v_n)$ se,

$$\lim_{n\to\infty} \frac{u_n}{v_n} = 0,$$

(iii) a sucessão u_n é assimptoticamente igual a v_n e escreve-se $u_n \sim v_n$ se,

$$\lim_{n\to\infty} \frac{u_n}{v_n} = 1.$$

Definição 2.1.3. Uma sucessão de funções $\{\phi_k(x)\}_{k\geq 0}$ diz-se uma sucessão assimptótica quando $x \to x_0$ se

$$\frac{\phi_{k+1}(x)}{\phi_k(x)} = o(1) \text{ quando } x \to x_0,\ k = 0, 1, 2, \ldots$$

Existem dois exemplos de sucessões assimptóticas que são relevantes para a aproximação de Padé:

(i) $\left\{(x - x_0)^k\right\}_{k\geq 0}$, para x_0 finito,

(ii) $\left\{x^{-k}\right\}_{k\geq 0}$, para x_0 infinito.

Definição 2.1.4. Diz-se que uma série formal $\sum_{k=0}^{\infty} c_k\phi_k(x)$ representa uma função $f(x)$ assimtoticamente quando $x \to x_0$ e escreve-se

$$f(x) \sim \sum_{k=0}^{\infty} c_k\phi_k(x) \text{ quando } x \to x_0,$$

se:

(i) $\{\phi_k(x)\}_{k\geq 0}$ é uma sucessão assimptótica,

(ii) para todo $n \geq 0$ tem-se,

$$f(x) - \sum_{k=0}^{n} c_k\phi_k(x) = \mathcal{O}\left(\phi_{n+1}(x)\right) \text{ quando } x \to x_0.$$

Proposição 2.1.2. Seja $\{\phi_k(x)\}_{k\geq 0}$ uma sucessão assimptótica. As seguintes afirmações são equivalentes:

(i)
$$f(x) \sim \sum_{k=0}^{\infty} c_k\phi_k(x) \text{ quando } x \to x_0.$$

(ii)
$$f(x) - \sum_{k=0}^{n} c_k\phi_k(x) = o\left(\phi_n(x)\right) \text{ quando } x \to x_0, \quad \text{para todo } n \geq 0.$$

(iii)
$$\lim_{x\to x_0} \frac{f(x) - \sum_{k=0}^{n} c_k\phi_k(x)}{\phi_{n+1}(x)} \text{ existe e é igual a } c_{n+1}, \quad \text{para todo } n \geq 0.$$

2.2 Aproximação de Padé de séries de potências

Começamos com a definição clássica de aproximação de Padé (AP), frequentemente chamada de *aproximação de Frobenius Padé* ou de *aproximação de Padé linear*.

2.2.1 Definições e Notações

Definição 2.2.1 (Aproximação de Padé linear)**.** Sejam p, q dois números inteiros não negativos e $f(z) = \sum_{k=0}^{\infty} c_k z^k$ uma série formal de potências na variável $z \in \mathbb{C}$. O aproximante de Padé linear de ordem (p, q) da série $f(z)$ é a função racional

$$\Phi_{p,q}(z) = \frac{N_{p,q}(z)}{D_{p,q}(z)}$$

com,

$$N_{p,q} \in \mathcal{P}_p,\ D_{p,q} \in \mathcal{P}_q,\ D_{p,q}(z) \not\equiv 0 \tag{2.1}$$

onde as séries de potências f e $fD_{p,q} - N_{p,q}$ coincidem até uma ordem ℓ tão alta quanto possível. Ou seja os polinómios $N_{p,q}$ e $D_{p,q}$ satisfazem as condições

$$\left(fD_{p,q} - N_{p,q}\right)(z) = \mathcal{O}\left(z^{\ell}\right) \text{ quando } z \to 0 \tag{2.2}$$

onde, dados os inteiros não negativos p e q, ℓ é a maior ordem para a qual as condições (2.1) e (2.2) se verificam.

O cálculo dos coeficientes do polinómio do denominador, $D_{p,q}(z) = \sum_{k=0}^{q} b_k z^k$ pode fazer-se resolvendo um sistema de q equações lineares homogéneo nas $q+1$ incógnitas $b_0, b_1, \ldots, b_q$. Usando a convenção, $c_i = 0$ se $i < 0$, pode-se escrever este sistema na forma matricial

$$\begin{bmatrix} c_{p+1} & c_p & \cdots & c_{p-q+1} \\ c_{p+2} & c_{p+1} & \cdots & c_{p-q+2} \\ \vdots & \vdots & & \vdots \\ c_{p+q} & c_{p+q-1} & \cdots & c_p \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.3}$$

Determinados os coeficientes de $D_{p,q}(z)$ determina-se o polinómio do numerador $N_{p,q}(z) = \sum_{k=0}^{q} a_k z^k$ usando a equação matricial

$$\begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} c_0 & 0 & \cdots & 0 \\ c_1 & c_0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ c_p & c_{p-1} & \cdots & c_{p-q} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_q \end{bmatrix} \tag{2.4}$$

deste modo, existe sempre um aproximante de Padé linear $\Phi_{p,q}(z)$. Por vezes, para enfatizar que o aproximante de Padé é relativo à série (ou à função) f, usaremos a notação alternativa $[p/q]_f(z)$. Geralmente o polinómio $D_{p,q}(z)$ (e consequentemente $N_{p,q}(z)$) não é univocamente determinado, dado que a característica da matriz do sistema (2.3) pode eventualmente ser menor do que q. Contudo verifica-se que a função racional $\Phi_{p,q}(z)$ é única [BGM96]. Se $D_{p,q}(0) \neq 0$ então $f(z) - \Phi_{p,q} = \mathcal{O}\left(z^{p+q+1}\right)$ quando $z \to 0$, e se $D_{p,q}(0) = 0$ tem-se $f(z) - \Phi_{p,q} = \mathcal{O}\left(z^{p+q+1-\omega_{p,q}}\right)$, quando $z \to 0$, onde $\omega_{p,q}$ é um inteiro positivo, chamado de *índice de deficiência*. Para evitar este problema, Baker [BJ73] usou a seguinte definição de aproximante de *Baker-Padé* ou aproximante de Padé *não linear*.

Definição 2.2.2 (Aproximação de Padé não linear)**.** Sejam p, q dois números inteiros não negativos e $f(z) = \sum_{k=0}^{\infty} c_k z^k$ uma série formal de potências. O aproximante de Padé não linear de ordem (p, q) da série $f(z)$ é, caso exista, a função racional $R_{p,q}(z) = \frac{N_{p,q}(z)}{D_{p,q}(z)}$, onde $N_{p,q}(z)$ e $D_{p,q}(z)$ são polinómios na variável z que satisfazem:

$$\mathrm{gr}\left(N_{p,q}\right) = p,\ \mathrm{gr}\left(D_{p,q}\right) = q,\ D_{p,q}(0) = 1 \tag{2.5}$$

$$f(z) - R_{p,q}(z) = \mathcal{O}\left(z^{p+q+1}\right) \text{ quando } z \to 0 \tag{2.6}$$

$R_{0,0}$	$R_{0,1}$	$R_{0,2}$	$R_{0,3}$	$\cdots$
$R_{1,0}$	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	$\cdots$
$R_{2,0}$	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	$\cdots$
$R_{0,3}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

Tabela 2.1: Tabela de Padé

onde $\mathrm{gr}\,(P)$ representa o grau de um polinómio P.

Da condição (2.6), conclui-se que os polinómios $N_{p,q}(z)$ e $D_{p,q}(z)$ também satisfazem a condição (2.2). Então, os aproximantes de Padé não lineares podem determinar-se usando as equações (2.3) e (2.4) fixando a condição $D_{p,q}(0) = 1$. Ou seja os aproximantes não lineares, se existirem, são determinados fazendo $b_0 = 1$ nas equações (2.3) e (2.4). Usando a conhecida regra de Cramer tem-se

$$R_{p,q}(z) = \frac{\begin{vmatrix} z^q S_{p-q}(z) & z^{q-1} S_{p-q+1}(z) & \cdots & z^0 S_p(z) \\ c_{p-q+1} & c_{p-q+2} & \cdots & c_{p+1} \\ c_{p-q+2} & c_{p-q+3} & \cdots & c_{p+2} \\ \vdots & \vdots & & \vdots \\ c_p & c_{p+1} & \cdots & c_{p+q} \end{vmatrix}}{\begin{vmatrix} z^q & z^{q-1} & \cdots & 1 \\ c_{p-q+1} & c_{p-q+2} & \cdots & c_{p+1} \\ c_{p-q+2} & c_{p-q+3} & \cdots & c_{p+2} \\ \vdots & \vdots & & \vdots \\ c_p & c_{p+1} & \cdots & c_{p+q} \end{vmatrix}}, \tag{2.7}$$

onde $S_\ell(z) = \sum_{k=0}^{\ell} c_k z^k$, $\ell = 0, 1, 2, \ldots$, e $c_\ell = 0$ se $\ell < 0$. Deste modo, podemos escrever $R_{p,q}(z)$ em função dos coeficientes $b_0 = 1, b_1, \cdots, b_q$ da forma

$$R_{p,q}(z) = \frac{\sum_{k=0}^{p} b_k z^k S_{p-k}(z)}{\sum_{k=0}^{q} b_k z^k}.$$

Finalmente, a existência do aproximante de Padé não linear não está garantida, contudo no caso de existir é único.

2.2.2 A tabela de Padé

Frequentemente organizam-se os aproximantes de Padé numa tabela, a qual é conhecida por *tabela de Padé*, ver Tabela 2.1.

Definição 2.2.3. Um aproximante de Padé $R_{p,q}$ diz-se *normal* se ocorre exatamente uma vez na tabela de Padé. Diz-se que a tabela de Padé é *normal* se todas as suas entradas forem normais.

A normalidade dos aproximantes de Padé está relacionada com a estrutura das entradas nulas da chamada *tabela-C* que iremos definir. De forma análoga à tabela de Padé, organizando os coeficientes $C_{p,q} = D_{p,q}(0)$, $p, q \geq 0$, obtém-se a *tabela-C*. Os coeficientes $C_{p,q}$ podem calcular-se usando o determinante que ocorre no denominador da relação (2.7). Ou seja tem-se

$$C_{p,0} = 1; \quad C_{p,q} = \begin{vmatrix} c_{p-q+1} & c_{p-q+2} & \cdots & c_p \\ c_{p-q+2} & c_{p-q+3} & \cdots & c_{p+1} \\ \vdots & \vdots & & \vdots \\ c_p & c_{p+1} & \cdots & c_{p+q-1} \end{vmatrix}, \quad p \geq 0, q \geq 1 \tag{2.8}$$

onde, $c_\ell = 0$ se $\ell < 0$.

Os seguintes resultados relacionam a estrutura da tabela de Padé com a estrutura das entradas nulas da tabela-C [BGM96].

Teorema 2.2.1. As seguintes afirmações são equivalentes:

1. $R_{p,q}$ é normal.
2. O numerador $N_{p,q}$ e o denominador $D_{p,q}$ de $R_{p,q}$ têm grau p e q, respetivamente, e $f(z) - R_{p,q} = \sum_{k=p+q+1}^{\infty} d_k z^k$, com $d_{p+q+1} \neq 0$.
3. Os determinantes $C_{p,q}$, $C_{p,q+1}$, $C_{p+1,q}$ e $C_{p+1,q+1}$ não se anulam.

Teorema 2.2.2. A tabela de Padé de uma série formal $f(z) \sim \sum_{k=0}^{\infty} c_k z^k$ é normal se e somente se $C_{p,q} \neq 0$, para todo $p, q \geq 0$.

Note-se que como $C_{p,1} = c_p$ então $c_p \neq 0$, para todo o inteiro não negativo, é uma condição necessária para que a tabela de Padé seja normal.

Teorema 2.2.3. Entradas nulas numa tabela-C ocorrem em blocos quadrados cercados por entradas não nulas, exceto num bloco infinito. Para um bloco, quadrado, de dimensão ℓ constituído pelas entradas nulas $C_{p,q}$, $m+1 \leq p \leq m+\ell$, $n+1 \leq q \leq n+\ell$ tem-se $R_{p,q} = R_{m+1,n+1}$ para $p \geq m$, $q \geq n$ e $p+q \leq m+n+\ell$ e os aproximantes de Padé de ordem (p, q) são iguais para $m \leq p \leq p+\ell$ e $n \leq q \leq n+\ell$.

2.3 Convergência de Aproximantes de Padé

Iremos apenas resumir alguns resultados relativos à convergência uniforme e à convergência em medida. De referir que alguns resultados além de estabelecerem a convergência de algumas sucessões de aproximantes de Padé também estabelecem aceleração de convergência.

No que se segue iremos considerar as seguintes sucessões de aproximantes de Padé:

1. *sucessões linha*, $\{R_{p,\ell}\}_{p\geq 0}$, com ℓ fixo,
2. *sucessões coluna*, $\{R_{\ell,q}\}_{q\geq 0}$, com ℓ fixo,
3. *sucessões para-diagonais*, $\{R_{q+\ell,q}\}_{q\geq 0}$, com ℓ fixo. Para $\ell = 0$ tem-se a *sucessão diagonal*,

2.3.1 Convergência uniforme

Convergência de funções meromorfas:

Um resultado de convergência de sucessões linha de AP para funções meromorfas foi originalmente demonstrado por Montessus de Ballore [dMdB02]. Este teorema além de estabelecer a convergência uniforme também estabelece a aceleração de convergência. Mais exatamente tem-se

Teorema 2.3.1 (de Montessus)**.** Seja $f(z)$ analítica em $z = 0$ e meromorfa em $\mathcal{B} = \{z : |z| < r\}$, $r > 0$, com ℓ pólos, $\eta_1, \eta_2, \ldots, \eta_\ell$, (contando com eventuais multiplicidades) em $\mathcal{B}$. Então a sucessão linha $\{R_{p,\ell}\}_{p\geq 0}$ converge uniformemente para $f(z)$ em todo o subconjunto compacto de $\mathcal{B} \setminus \{\eta_1, \eta_2, \ldots, \eta_\ell\}$, e,

$$\limsup_{p\to\infty} |f(z) - R_{p,\ell}(z)|^{1/p} \leq \frac{|z|}{r}. \tag{2.9}$$

O seguinte resultado, [GH66], é relativo aos pólos dos aproximantes de Padé de funções meromorfas nas condições do teorema, e generaliza o teorema de König, [Kön84].

Teorema 2.3.2. Seja $f(z)$ analítica em $z = 0$ e meromorfa em $\mathcal{B} = \{z : |z| < r\}$, com ℓ pólos $\eta_1, \eta_2, \ldots, \eta_\ell$ tais que $|\eta_m| \geq |\eta_k|$, $k = 1, 2, \ldots \ell$. Definindo o polinómio $D(z) = \prod_{k=1}^{\ell} (1 - z/\eta_k)$ tem-se

$$\limsup_{p\to\infty} |D(z) - D_{p,\ell}(z)|^{1/p} \leq \frac{|\eta_m|}{r}, \tag{2.10}$$

onde $D_{p,q}(z)$ é o denominador de $R_{p,q}(z)$.

Convergência de séries de momentos

Enquanto os resultados de convergência de AP de funções meromorfas se referem a sucessões de AP linha, os resultados sobre a convergência de AP de séries de momentos são relativos a sucessões para-diagonais. Existem na literatura várias definições de funções de Stieltjes, e, consequentemente várias definições de séries de Stieltjes. Iremos, nesta secção, adotar as seguintes definições dadas em [BGM96] e [Sid03].

Definição 2.3.1. Sejam $I =]a, b[\subset \mathbb{R}$ e $\sigma(t)$ uma função real não decrescente e com um número infinito de pontos de crescimento no intervalo I. Considere-se a função $\hat{\sigma}(z)$ definida pelo integral de Stieltjes,

$$\hat{\sigma}(z) = \int_I \frac{d\sigma(t)}{1+tz}. \tag{2.11}$$

Se $0 \leq a < b \leq +\infty$, $\hat{\sigma}(z)$ diz-se uma *função de Stieltjes* e se $-\infty \leq a < 0 < b \leq +\infty$ então, a função $\hat{\sigma}(z)$ diz-se *de Hamburger*.

Se existirem os momentos, m_k, de $\sigma(t)$

$$m_k = \int_I t^k d\sigma(t), \ k = 0, 1, 2, \ldots, \tag{2.12}$$

então, a série formal $\sum_{k=0}^{\infty} m_k(-z)^k$ diz-se a série de momentos associada a $\sigma(t)$. No caso de $\hat{\sigma}(z)$ ser uma função de Stieltjes (de Hamburger) diz-se que $\sum_{k=0}^{\infty} m_k(-z)^k$ é uma série de Stieltjes (de Hamburger).

Estas funções verificam as seguintes propriedades:

Teorema 2.3.3.

1. Se $\hat{\sigma}$ é uma função de Stieltjes então, é holomorfa em $\mathbb{C}\backslash]-1/a, -1/b[$ e $[\hat{\sigma}(z)]^* = \hat{\sigma}(z^*)$,

2. Se $\hat{\sigma}$ é uma função de Hamburger então, é holomorfa em $\mathbb{C}\backslash]-\infty, -1/b] \cup [-1/a, +\infty[$ e $[\hat{\sigma}(z)]^* = \hat{\sigma}(z^*)$,

3. Se $\sigma(t)$ tiver momentos finitos m_k, $k = 0, 1, 2, \ldots$, então,

$$\hat{\sigma}(z) \sim \sum_{k=0}^{\infty} m_k(-z)^k, \quad (z \to 0)$$

   e além disso, se $]a, b[$ for limitado então a série dos momentos tem raio de convergência r, $0 < r < \infty$, e se $a = -\infty$ ou $b = +\infty$ então a série dos momentos tem raio de convergência $r = 0$.

Apenas iremos referir os resultados relativos às series de Stieltjes. Começamos com um resultado relativo à localização dos pólos de sucessões para-diagonais.

Teorema 2.3.4. Os aproximantes de Padé, $R_{p+\ell,p}$, com $\ell \geq -1$, de uma série de Stieltjes são normais e possuem pólos simples. Estes pólos situam-se na semi-reta $\mathbb{R}^- = \{x \in \mathbb{R} : \ x < 0\}$ e têm resíduos positivos.

Os resultados de convergência de séries de Stieltjes dependem de duas situações. Na primeira o intervalo $I = [a, b]$, $a \geq 0$, é limitado, (caso em que a série de Stieltjes tem raio de convergência positivo) e na segunda situação o intervalo I é ilimitado ($I =]a, +\infty[$) (neste caso, a série de Stieltjes é apenas formal). O resultado seguinte resume as duas situações

Teorema 2.3.5. Seja $\{R_{p+\ell,p}\}_{p\geq 0}$, uma sucessão para-diagonal de aproximantes de Padé, com $\ell \geq -1$, de uma série de Stieltjes, $\sum_{k=0}^{\infty} m_k(-z)^k$, onde $m_k = \int_I t^k d\sigma(t)$, $k = 0, 1, 2, \ldots$. Então,

1. se $0 \leq a < b < +\infty$ a sucessão converge para $\hat{\sigma}(z)$ em $\mathcal{D} = \mathbb{C}\backslash]-\infty, -1/b]$ a uma taxa pontual

$$\limsup_{p\to+\infty} |\hat{\sigma}(z) - R_{p+\ell,p}(z)|^{1/n} \leq \left| \frac{\sqrt{1+bz}-1}{\sqrt{1+bz}+1} \right| < 1$$

   (com a convenção que $\mathrm{Arg}(\sqrt{1+bz}) > 0$ se $z > -1/b$). A convergência é uniforme em todo o compacto $\mathcal{K} \subset \mathcal{D}$.

2. se $I =]0, \infty[$ e se a série de Stieltjes satisfizer a condição de Carlman (a série $\sum_{k\geq 1}(m_k)^{-1/(2k)}$ diverge) então a sucessão para-diagonal converge uniformemente para $\hat{\sigma}(z)$ em $\mathcal{D}_{r,\epsilon}$, onde

$$\mathcal{D}_{r,\epsilon} = \left\{ z \in \mathbb{C} \ : \ |z| \leq r \text{ e } \mathrm{dist}\left(z, \mathbb{R}_0^-\right) \geq \epsilon \right\},$$

   para todo $0 < \epsilon < r$.

2.3.2 Convergência em medida

Observamos que os resultados sobre convergência em medida de sucessões de AP não garantem a convergência pontual em nenhum ponto do plano complexo. Na realidade os resultados que iremos expor são válidos no plano complexo exceto num conjunto de medida arbitrariamente pequena. Iremos sempre supor a existência de todos os AP mencionados. Começamos com uma versão fraca do teorema de Montessus 2.3.1.

Teorema 2.3.6. Seja $f(z)$ analítica em $z = 0$ e meromorfa em $\mathcal{B} = \{z \ : \ |z| < r\}$, com ℓ pólos, contando com eventuais multiplicidades, em $\mathcal{B}$. Considere uma sucessão coluna de AP $\{R_{p,q}\}_{p\geq 0}$ da função f, com $q \geq \ell$. Dados dois números reais positivos ϵ e δ arbitrariamente pequenos, existe ℓ_{min} tal que $|f(z) - R_{p,q}(z)| < \epsilon$ para todo $q \geq \ell_{min}$ e para todo $z \in \mathcal{B} \setminus \mathcal{E}_\ell$, onde $\mathcal{E}_\ell \subset \mathbb{C}$ é um conjunto com medida inferior a δ.

Os seguintes corolários, do teorema 2.3.6, generalizam a convergência em medida a outras sucessões de AP.

Corolário 2.3.1. Sob as hipóteses do teorema 2.3.6 tem-se $|f(z) - R_{p_k,q_k}| < \epsilon$, para todo $k > k_0$ e para todo $z \in \mathcal{B} \setminus \mathcal{E}_k$, onde $\mathcal{E}_k$ é um conjunto com medida inferior a δ desde que:

i) $\lim_{k\to\infty} \frac{p_k}{q_k} = \infty$, $(q_k \neq 0)$,

ii) $q_k \geq \ell$, para todo $k > k_0$.

Corolário 2.3.2. Sob as hipóteses do teorema 2.3.6 existe um $k > \ell_{min}$ tal que $|f(z) - R_{p,q}| < \epsilon$ para todo $z \in \mathcal{B} \setminus \mathcal{E}_k$, onde $\mathcal{E}_k$ é um conjunto com medida inferior a δ.

O seguinte resultado abrange o teorema 2.3.6 e os seus dois corolários.

Teorema 2.3.7. Seja $f(z)$ analítica em $z = 0$ e também analítica em $\mathcal{B} = \{z : |z| < R\}$, com ℓ pólos (contando com eventuais multiplicidades) em $\mathcal{B}$. Considere uma sucessão R_{p_k,ℓ_k} de AP de f com $\ell_k \geq \ell$ e $\lim_{k\to\infty} \frac{p_k}{\ell_k} = \infty$. Dados dois números positivos arbitrariamente pequenos ϵ e δ existe k_0 tal que $|f(z) - R_{p_k,\ell_k}| < \epsilon$ para todo $k > k_0$ e para todo o $z \in \mathcal{B} \setminus \mathcal{E}_k$ onde $\mathcal{E}_k$ é um conjunto com medida inferior a δ.

Existem igualmente resultados, sobre convergência em medida, para outros tipos de sucessões tais como: sucessões raio e sucessões diagonais. Os resultados relativos a sucessões diagonais foram inicialmente estabelecidos por Nuttall [Nut70] e posteriormente estendidos a sucessões raio por Pommerenke [Pom73] e outros autores. Apenas iremos mencionar dois resultados, de interesse para este trabalho, sendo que o primeiro foi estabelecido por Nuttall [Nut70] e o segundo por Gammel e Nuttall [GN73].

Teorema 2.3.8 (Nuttall)**.** Seja f uma função meromorfa. Dados dois números positivos ϵ e δ, então existe p_0 tal que, para todo $p > p_0$ tem-se $|f(z) - R_{p,p}(z)| < \epsilon$, em todo o conjunto $\mathcal{K} \setminus \mathcal{E}_p$, onde $\mathcal{K} \subset \mathbb{C}$ é um compacto e $\mathcal{E}_p$ um conjunto com medida inferior a δ.

O resultado seguinte é relativo a funções que pertencem à classe $\mathcal{A}$ (de Borel) de funções *quase analíticas*. Mais exatamente, à classe das funções representadas por uma série da forma

$$f(z) = \sum_{k=0}^{\infty} \frac{A_k}{1 - z\alpha_k}, \quad A_k, \alpha_k \in \mathbb{C} \tag{2.13}$$

onde os coeficientes A_k satisfazem a condição

$$|A_k| < C \exp(-k^{1+\epsilon_1}), \quad \text{com } \epsilon_1 > 0. \tag{2.14}$$

Observações:

i *Uma função f representada por uma série da forma (2.13) com os coeficientes A_k tais que a série*

$$\sum_{k\geq 0} \frac{\log\log(1/|A_k|)}{\log(1/|A_k|)}, \quad \textit{converge} \tag{2.15}$$

é quase analítica [BJ17].

ii *Carleman [Car26] mostrou que se os coeficientes A_k satisfizerem a condição (2.14) então, f é uma função quase analítica.*

iii *Se escolhermos os elementos do conjunto $\{\alpha_k\}_{k\geq 0}$ de forma a estarem densamente distribuídos numa circunferência $|z| = r$, $r > 0$, e se os A_k satisfizerem a condição (2.14) então a função é quase analítica e a sua série de potências deduzida de (2.13) tem uma fronteira natural na circunferência $|z| = r^{-1}$.*

Teorema 2.3.9 (Gammel-Nuttall)**.** Seja f uma função em $\mathcal{A}$ com, $|\alpha_k| = 1$, $k = 0, 1, \ldots$. Dados números positivos ϵ e δ arbitrariamente pequenos e um número inteiro ℓ e seja $R_{p+\ell,p}$ onde $p + \ell > 0$ uma sucessão para-diagonal de AP de f, seja ainda, $\mathcal{K} \subset \mathbb{C}$ uma região compacta arbitrária então, existe um número natural p_0 tal que

$$|f(z) - R_{p+\ell,p}(z)| < \epsilon$$

para todo o $p > p_0$ e para todo o $z \in \mathcal{K} \setminus \mathcal{E}$, onde o conjunto $\mathcal{E}$ tem medida inferior a δ.

Note-se que este resultado garante que toda a sucessão para-diagonal de AP converge em medida em qualquer compacto no plano complexo para a função f. Isto significa que se escolhermos um conjunto de coeficientes $\{\alpha_k\}_{k\geq 0}$ densamente distribuídos na circunferência unitária então, a função f tem fronteira natural na circunferência unitária e a convergência em medida para f verifica-se em todo o compacto $\mathcal{K} \setminus \mathcal{E}$, onde o conjunto $\mathcal{E}$ contém a fronteira natural de f.

2.4 Estimação de singularidades

Para certas famílias de funções, os pólos de sucessões de AP determinados por uma série de potências $f(z) = \sum_{k\geq 0} c_k z^k$ tendem para as singularidades da função f. Tendo em vista a localização de singularidades de funções dadas por séries de potências via determinação de pólos de sucessões de aproximantes de Padé, estamos interessados no chamado *problema inverso*. Ou seja, supondo que os pólos de uma sucessão de AP de uma função f tendem para elementos num conjunto Λ pode-se concluir que a função f é singular em Λ? Um resultado chave relativo a este problema é o seguinte, [Fab96, Die57],

Teorema 2.4.1 (Fabry)**.** Seja $f(z) \sim \sum_{k\geq 0} c_k z^k$ uma série de potências, tal que o limite $\lim_{k\to\infty} \frac{c_k}{c_{k+1}}$ existe e é igual a $\lambda \neq 0$. Então a série converge uniformemente em todo o compacto $\mathcal{K} \subset \{|z| < |\lambda|\})$ e λ é um ponto singular da função f.

Se $c_{p+1} \neq 0$, $p = 0, 1, \ldots$, então os pólos dos AP da segunda coluna da tabela de Padé, $x_{p,1}$, $p = 0, 1, \ldots$, são precisamente $x_{p,1} = \frac{c_p}{c_{p+1}}$ e tem-se o seguinte corolário do teorema 2.4.1

Corolário 2.4.1. Dada uma série de potências $\sum_{k\geq 0} c_k z^k$. Se a sucessão dos pólos $\{x_{p,1}\}_{p\geq 0}$ tende para $\lambda \neq 0$ então, $f(z) = \sum_{k\geq 0} c_k z^k$ é analítica no disco $\mathcal{D}_\lambda = \{|z| < |\lambda|\}$ e λ é uma singularidade da função f.

O seguinte resultado generaliza os resultados obtidos por Fabry a outras sucessões coluna de AP, [VGP81] e [Sue85].

Teorema 2.4.2 (Suetin)**.** Seja $\sum_{k\geq 0} c_k z^k$ uma série de potências cujos coeficientes satisfazem as condições: para todo $q \in \mathbb{N}$ fixo e para todo $p \in \mathbb{N}$ suficientemente grande o AP $R_{p,q}$ tem precisamente q pólos finitos $x_{p,1}, x_{p,2}, \ldots, x_{p,q}$ e os limites

$$\lim_{p\to\infty} x_{p,k} = \lambda_k, \quad k = 1, 2, \ldots, q.$$

Então:

(i) $\sum_{k\geq 0} c_k z^k$ converge uniformemente no disco $\mathcal{D}_m = \{|z| < \min_{1\leq i\leq q} |\lambda_i|\}$;

(ii) a função $f(z) = \sum_{k\geq 0} c_k z^k$ admite uma continuação meromórfica ao disco $\mathcal{D}_M = \{|z| < \max_{1\leq i\leq q} |\lambda_i|\}$;

(iii) a função f tem no máximo $q-1$ pólos no disco $\mathcal{D}_M$ e todos os pontos λ_i, $i = 1, 2, \ldots, q$ são singularidades de f,

(iv) as singularidades $\lambda_i \in \mathcal{D}_M$ são pólos e a função f não possui outros pólos em $\mathcal{D}_M$.

Frequentemente, em aplicações práticas, não temos acesso aos coeficientes exatos de séries de potências. Consequentemente, os AP são calculados a partir de coeficientes aproximados. Deste modo, a qualidade das AP irá ser influenciada pelos erros nestas aproximações. Um fator que influencia drasticamente a qualidade da aproximação de um AP é a localização dos seus pólos. Estas observações dão o mote à secção seguinte, que se baseia fundamentalmente nos resultados de experiências numéricas obtidos por M. Froissart.

2.5 Localização de pólos e zeros de AP de séries de potências perturbadas

Tendo em vista estudar o efeito causado nos aproximantes de Padé devido aos erros nos coeficientes de séries de potências, M. Froissart [Fro69], simulou os erros nos coeficientes, $c_k,\ k = 0, 1, \ldots$, usando séries com coeficientes aleatórios (a que chamaremos "ruídos"). Iremos reproduzir algumas das experiências numéricas efetuadas por Froissart, bem como algumas contribuições posteriores para explicar os resultados obtidos nessas experiências. Para o efeito iremos considerar dois tipos de ruídos, designados por: ruídos do tipo I e do tipo II e definidos por:

$$T_\epsilon(z) = \sum_{k=0}^{\infty} \epsilon r_k z^k, \quad \text{ruído do tipo I,} \tag{2.16}$$

$$T_\omega(z) = \sum_{k=0}^{\infty} \omega \frac{r_k}{2^k} z^k \quad \text{ruído do tipo II} \tag{2.17}$$

onde ϵ e ω são números reais positivos e os r_k são números complexos aleatórios uniformemente distribuídos no disco unitário, ou seja $|r_k| \leq 1$, k=0,1,....

Exemplo 2.5.1. Froissart considerou as funções $f(z) = 1/(1-z)$ e $g(z) = \ln(1-z)$ e representadas, respetivamente, pelas séries de Taylor, centradas em $z = 0$, $S_f(z) = \sum_{k=0}^{\infty} z^k$ e $S_g(z) = -\sum_{k=0}^{\infty} \frac{z^k}{k}$. Notamos que:

- a função f é uma função racional, com um pólo em $z_1 = 1$, logo (devido à propriedade de consistência dos AP) tem-se $[p/q]_f = f$, para todos os inteiros p e q tais que $p \geq 0$ e $q \geq 1$ (ou seja, a tabela de Padé tem um bloco infinito). Além disso, a série $S_f(z)$ é absolutamente convergente em $|z| < 1$.

- a função g tem: um zero no ponto $z = 0$, dois pontos de ramificação, em $z = 1$ e $z = \infty$, e um corte de ramificação unindo os pontos de ramificação. A série $S_g(z)$ é absolutamente convergente em $|z| < 1$.

Froissart perturbou as séries $S_f(z)$ e $S_g(z)$ com ruídos do tipo I e II, $S_{f_\epsilon}(z) = \sum_{k=0}^{\infty}(1 + \epsilon r_k)z^k$, $S_{f_\omega}(z) = \sum_{k=0}^{\infty}(1 + \epsilon\frac{r_k}{2^k})z^k$, $S_{g_\epsilon}(z) = \sum_{k=0}^{\infty}(-\frac{z^k}{k} + \omega r_k)z^k$ e $S_{g_\omega}(z) = \sum_{k=0}^{\infty}(-\frac{z^k}{k} + \omega\frac{r_k}{2^k})z^k$, e estudou a localização dos pólos e zeros de AP diagonais destas séries.

Para os aproximantes $[p/p]_{S_{f_\epsilon}}$ observou os seguintes comportamentos:

(1) para cada $p \geq 1$, $[p/p]_{S_{f_\epsilon}}$ possui um pólo estável ξ_1 perto da singularidade de f, $z_1 = 1$. Este pólo considera-se estável no sentido que se tem, $|z_1 - \xi_1| = \mathcal{O}(\epsilon)$, e esta distância diminui com o crescimento do valor de p.

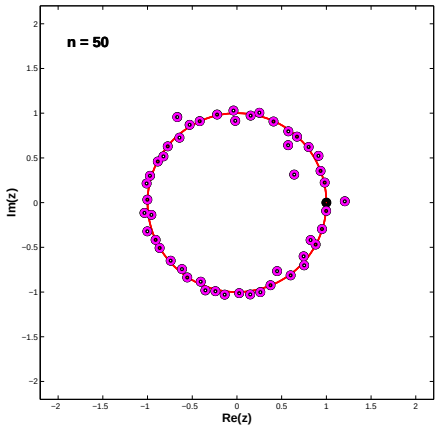


Figura 2.1: Localização dos pólos (pontos a preto) e zeros (circunferências a magenta) do AP diagonal, $n = 50$, da série S_{f_ϵ}, com $\epsilon = 10^{-3}$. Não indicamos o zero fantasma na Figura para facilitar a leitura da mesma.

(2) existe um zero instável η_1, no sentido que a sua localização depende do valor de n, tal que $|\eta_1| = \mathcal{O}(\epsilon^{-1})$, para $n = 1$, e $|\eta_1|$ aumenta com o crescimento de n. O comportamento da localização deste zero, usualmente chamado de "zero fantasma", reflete o facto de a função f restrita à recta real tender para zero quando $x \to \infty$.

(3) para valores de $n > 1$ os zeros e pólos restantes η_k e ξ_k, $k = 2, \ldots, n$, respetivamente, surgem agrupados aos pares. Mais precisamente, a distância $|\xi_k - \eta_k|$ é pequena, para $n = 1$, e diminui quando o valor de n cresce. Além disso, estes pares localizam-se perto da circunferência de raio unitário.

Este comportamento é ilustrado na Figura 2.1, onde se apresenta o conjunto dos pólos e dos zeros de $[50/50]_{S_{f_\epsilon}}$ junto com a circunferência de raio unitário. Estes resultados correspondem a uma realização aleatória de ruído do tipo I, como em (2.16), com $\epsilon = 10^{-3}$. Neste caso o zero fantasma satisfaz a condição $|\eta_1| = \mathcal{O}(10^3)$ e por uma questão de escala não é mostrada na Figura.

Notamos que a ocorrência dos zeros e dos pólos agrupados aos pares, em que a distância entre dois elementos de cada par (η_k, ξ_k), $k = 2, \ldots, n$, é pequena e diminui com o aumento de n, traduz o facto acima referido, de que a tabela de Padé dos AP da função f contém um bloco infinito. Note-se que cada par contribui com fatores do numerador e do denominador do AP que se cancelam assimptoticamente. Informalmente, designa-se por *par de Froissart*, a um par constituído por um pólo e um zero (η, ξ) que têm esta propriedade, a distância entre eles é pequena e diminui quando o valor de n aumenta. Se um dado par de Froissart não representa a estrutura das singularidades da função f diremos que é um par de Froissart artificial, e caso ele represente de certa forma a estrutura das

singularidades da função f denominamo-lo de par de Froissart "genuíno". Formalmente, Stahl [Sta98] definiu um par de Froissart "artificial" como sendo um par (η, ξ) onde ξ é um *pólo espúrio* e η um zero muito próximo de ξ e definiu um par de Froissart "genuíno" como sendo um par (η, ξ) satisfazendo as mesmas condições assimptóticas dos pares artificiais com a exceção de que o pólo η não é espúrio. Contudo, a definição de Stahl não é útil em aplicações numéricas dado que a distinção entre pares de Froissart artificiais e genuínos dependem da definição de pólo espúrio, a qual é dada de uma forma assimptótica.

Para explicar a localização dos pares de Froissart dos AP diagonais da função perturbada S_{f_ϵ}, iremos abordar o seguinte resultado sobre séries de potências com coeficientes aleatórios [GTV87, Kah93].

Teorema 2.5.1. Sejam $(\Omega, \mathcal{A}, \mathcal{P})$ um espaço de probabilidade e X_n, $n = 0, 1, \ldots$, variáveis aleatórias complexas independentes, então dado $\omega \in \Omega$, o raio de convergência da série $\sum_{n=0}^{\infty} X_n(\omega) z^n$ é, com probabilidade um,

$$r(\omega) = \left(\limsup_{n \to \infty} |X_n(\omega)|^{1/n} \right)^{-1},$$

e além disso se, $0 < r(\omega) < \infty$ e se X_n forem simétricas (X_n é simétrica, se tiver a mesma distribuição que $-X_n$) então, $|z| = r(w)$ é uma fronteira natural, com probabilidade um, da função

$$F(z; \omega) = \sum_{n=0}^{\infty} X_n(\omega) z^n.$$

Resulta imediatamente deste teorema o seguinte

Corolário 2.5.1. Os ruídos do tipo I e do tipo II possuem fronteiras naturais, com probabilidade um, nas circunferências $|z| = 1$ e $|z| = 2$.

> **Observação:** *Deste modo, os pares de Froissart, localizados na vizinhança da circunferência de raio unitário, representam a fronteira natural do ruído do tipo* I.

Em seguida, iremos descrever os resultados obtidos na segunda experiência de Froissart, com a série S_{g_ω}.

Exemplo 2.5.2. Para a série perturbada $S_{g_\omega}(z)$ da função $g(z)$ observamos o seguinte comportamento dos zeros e pólos dos AP diagonais:

(1) para cada $n \geq 1$, $[n, n]_{S_{g_\omega}(z)}$ tem um zero estável η_1 próximo do zero da função g, $z = 0$. Para $n = 1$ tem-se $|\eta_1| = \mathcal{O}(\omega)$ e, $|\eta_1|$ diminui com o aumento de n.

(2) existe um pólo estável ξ_1 próximo do ponto de ramificação $z_1 = 1$. Para $n = 1$ tem-se $|\xi_1 - z_1| = \mathcal{O}(\omega)$ e, $|\eta_1 - z_1|$ diminui com o aumento de n.

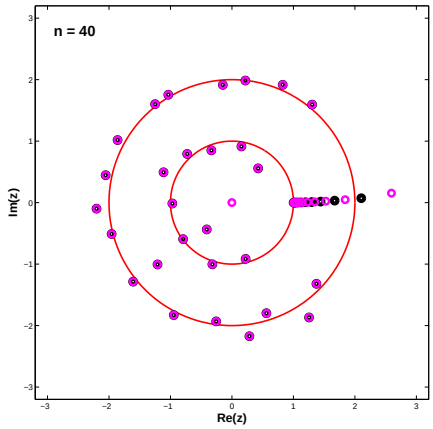


Figura 2.2: Distribuição dos zeros e pólos do AP diagonal da série S_{g_ω}, $n = 40$ e $\omega = 10^{-3}$. Os pólos estão assinalados com pontos a preto e os zeros com circunferências a magenta. Para facilitar a leitura da figura não incluímos todos os pólos e zeros que representam o corte da função g.

(3) existe um conjunto de pólos e zeros intercalados, os quais representam o corte de bifurcação da função g. Além disso, os elementos deste conjunto tendem a acumular-se próximo do ponto de ramificação $z_1 = 1$.

(4) os pólos e zeros restantes são pares de Froissart. Enquanto uns estão agrupados próximos da fronteira natural do ruído, a circunferência $|z| = 2$, os outros agrupam-se próximo da circunferência $|z| = 1$ e são originados pelo facto da função g possuir uma singularidade no interior da fronteira natural do ruído.

Este comportamento é ilustrado na Figura 2.2, onde a circunferência maior tem raio dois e representa a fronteira natural do ruído enquanto a circunferência menor, de raio um, representa a fronteira do disco de convergência da série $S_g(z)$ que representa a função g.

Observação:

Os pares de Froissart artificiais podem não ser apenas gerados por ruídos nos coeficientes da série. O método usado para calcular o AP pode igualmente gerar pares de Froissart artificiais. Por exemplo, nos testes que apresentámos as matrizes do sistema (2.3) *são mal condicionadas. Este efeito nota-se sobretudo no segundo exemplo onde existem pares de Froissart localizados próximo da fronteira natural do ruído artificialmente introduzido do tipo* II *e existem pares de Froissart, gerados pelo mau condicionamento da matriz do sistema* (2.3)*, localizados próximo da circunferência $z = 1$. No primeiro exemplo os pares de Froissart gerados pelo mau condicionamento da matriz do sistema*

(2.3) *não se distinguem dos pares de Froissart gerados pelo ruído artificialmente introduzido do tipo* I. *Estes factos permitem conjeturar que o ruído introduzido pelo mau condicionamento da matriz do sistema* (2.3) *é um ruído do tipo* I.

Estes problemas irão ser abordados com mais detalhe na secção 2.7, onde é abordada a localização de pólos e zeros de AP de séries ortogonais perturbadas.

2.6 Aproximantes de Padé de séries de polinómios ortogonais

O conceito de AP de funções representadas por séries de potências pode ser generalizado a funções representadas em séries ortogonais. Tal como na secção anterior, existem fundamentalmente duas classes de aproximantes de funções representadas em séries ortogonais. Os *aproximantes não lineares* e os *aproximantes de Frobenius-Padé* ou *aproximantes lineares*. Contudo, enquanto no caso de AP de séries de potências os dois conceitos quase que coincidem, no caso dos AP de séries ortogonais temos que distinguir os aproximantes lineares dos não lineares. Iremos, nesta secção, resumir os conceitos e resultados destes aproximantes dando especial ênfase aos aproximantes de Padé de séries de Chebyshev (ACP) e aos aproximantes de Padé de séries de Legendre (ALP).

2.6.1 Definições e Notações

Seja $f \in L^2_w(I)$, (B.2), uma função representada por uma série ortogonal

$$f(z) = \sum_{k=0}^{\infty} f_k \varphi_k(z), \tag{2.18}$$

onde $\{\varphi_i\}_{i\geq 0}$ é uma família de polinómios ortogonais num intervalo $I \subset \mathbb{R}$ relativamente a um produto interno

$$(u, v)_w = \int_I uvw \mathrm{d}z,$$

sendo $w(z) \geq 0$ uma função peso. Ou seja, definindo as funcionais

$$\mathcal{U}_k(f) = \frac{1}{\|\varphi_k\|_w^2} (f, \varphi_k)_w, \quad k \geq 0$$

tem-se

$$f_k = \mathcal{U}_k(f), \quad k \geq 0. \tag{2.19}$$

Definição 2.6.1. (Aproximante não linear) Dados dois inteiros não negativos p e q, define-se um aproximante não linear de ordem (p,q) da série (2.18) como sendo uma função racional $R_{p,q}(z) = N_{p,q}(z)/D_{p,q}(z)$ onde: $N_{p,q}(z) = \sum_{i=0}^{p} a_i\varphi_i(z)$, $D_{p,q}(z) = \sum_{i=0}^{q} b_i\varphi_i(z)$, e cuja expansão ortogonal satisfaz a condição

$$f(z) - R_{p,q}(z) = \sum_{k\geq p+q+1} d_k\varphi_k(z)$$

Logo, um aproximante não linear é determinado pelo sistema de equações não lineares cujas incógnitas são os coeficientes de $N_{p,q}$ e $D_{p,q}$,

$$\mathcal{U}_k(R_{p,q}) = f_k \quad k = 0, 1, \ldots p+q. \tag{2.20}$$

Definição 2.6.2. (Aproximante linear) Dados dois inteiros não negativos p e q, define-se um aproximante linear de ordem (p,q) da série (2.18) como sendo uma função racional

$$\Phi_{p,q}(z) = \frac{N_{p,q}(z)}{D_{p,q}(z)} \equiv \frac{\sum_{i=0}^{p} a_i\varphi_i(z)}{\sum_{i=0}^{q} b_i\varphi_i(z)}$$

que satisfaz

$$\mathcal{U}_k\left(D_{p,q}f - N_{p,q}\right) = 0 \quad k = 0, 1, \ldots, p+q. \tag{2.21}$$

Ou seja, exige-se que $D_{p,q}f - N_{p,q}$ seja ortogonal a $\varphi_k, k = 0, 1, \ldots$. Por vezes, quando pretendermos enfatizar que o aproximante de Padé linear (não linear) é relativo a uma função (série) f usaremos a notação $[p/q]_f$ ($\langle p/q\rangle_f$) em vez de $\Phi_{p,q}$ ($R_{p,q}$).

Enquanto para se determinar um aproximante não linear $R_{p,q}$ apenas é necessário o conhecimento dos coeficientes f_k, $k \leq p+q+1$, para se determinar um aproximante de Padé linear $\Phi_{p,q}$ necessita-se dos coeficientes f_k, $k \leq p+2q+1$. Portanto, os aproximantes lineares possuem à partida uma desvantagem relativamente aos aproximantes não lineares. Contudo, enquanto as condições (2.21) que determinam um aproximante linear formam um sistema de equações lineares as condições (2.20), que determinam um aproximante não linear formam um sistema de equações não lineares o que pode colocar problemas de existência ou dificultar o cálculo dos aproximantes não lineares.

2.6.2 Cálculo de AP lineares

Para se determinar aproximações Padé lineares, segue-se de perto os algoritmos indicados em [Mat01] e [Mat03].

Dado que $\{\varphi_k(z)\}_{k\geq 0}$ é uma família de polinómios ortogonais, tem-se que as condições (2.21) implicam que os coeficientes a_i, $i = 0, 1, \ldots, p$ e b_i, $i = 0, 1, \ldots, q$, satisfazem a equação

$$\sum_{k=0}^{q} b_k\varphi_k(z)f(z) - \sum_{k=0}^{p} a_k\varphi_k(z) = \sum_{k\geq p+q+1} e_k\varphi_k(z). \tag{2.22}$$

Definindo

$$\varphi_k(z)f(z) = \sum_{j\geq 0} h_{j,k}\varphi_j(z), \quad \text{onde } h_{j,k} = \mathcal{U}_j\left(\varphi_k f\right).$$

Agrupando os coeficientes homólogos tem-se o seguinte sistema de equações lineares equivalente

$$\sum_{i=0}^{q} h_{j,i}b_i = a_j, \quad j = 0, \ldots, p \tag{2.23}$$

$$\sum_{i=0}^{q} h_{j,i}b_i = 0, \quad j = p+1, \ldots, p+q. \tag{2.24}$$

As equações (2.24) formam um sistema de $q+1$ equações a q incógnitas. Definindo a matriz

$$\mathbf{H_{p,q}} = \begin{bmatrix} h_{p+1,0} & \cdots & h_{p+1,q-1} \\ \vdots & & \vdots \\ h_{p+q,0} & \cdots & h_{p+q,q-1} \end{bmatrix}, \tag{2.25}$$

tem-se [Mat01]

Proposição 2.6.1. Se $\det\left(\mathbf{H_{p,q}}\right) \neq 0$ então existe um e somente um AP linear

$$\Phi_{p,q}(z) = \frac{\sum_{i=0}^{p} a_i\varphi_i(z)}{\sum_{i=0}^{q} b_i\varphi_i(z)}$$

tal que $b_q = 1$ e os coeficientes a_p e e_{p+q+1} (o primeiro coeficiente do erro (2.22)) são não nulos.

Deste modo, usando a normalização $b_q = 1$, o AP é determinado na forma matricial por

$$\mathbf{H_{p,q}} \cdot \mathbf{b_{p,q}} = -\mathbf{h}_{p,q} \tag{2.26}$$

$$\mathbf{a_{p,q}} = \mathbf{G_{p,q}} \cdot \mathbf{b_{p,q}} + \mathbf{g_{p,q}} \tag{2.27}$$

onde,

$$\mathbf{a_{p,q}} = \begin{bmatrix} a_0 \ldots a_p \end{bmatrix}^T, \quad \mathbf{b_{p,q}} = \begin{bmatrix} b_0 \ldots b_{q-1} \end{bmatrix}^T,$$
$$\mathbf{g_{p,q}} = \begin{bmatrix} h_{0,q} \ldots h_{p,q} \end{bmatrix}^T, \quad \mathbf{h_{p,q}} = \begin{bmatrix} h_{p+1,q} \ldots h_{p+q,q} \end{bmatrix}^T,$$

e

$$\mathbf{G_{p,q}} = \begin{bmatrix} h_{0,0} & \cdots & h_{0,q-1} \\ \vdots & & \vdots \\ h_{p,0} & \cdots & h_{p,q-1} \end{bmatrix}.$$

Para calcular as entradas destas matrizes, geralmente não é viável determinar os coeficientes $h_{j,i}$ calculando os integrais envolvidos na sua definição (2.19). Contudo os coeficientes $h_{j,i}$ satisfazem uma relação de recorrência. Dada uma família de polinómios ortogonais $\{\varphi_i\}_{i\geq 0}$ relativamente a uma função peso w e com relação de recorrência associada

$$x\varphi_i(x) = \alpha_i\varphi_{i+1}(x) + \beta_i\varphi_i(x) + \gamma_i\varphi_{i-1}(x), \quad i \geq 0 \tag{2.28}$$

os coeficientes $h_{i,j}$ satisfazem a relação de recorrência [Mat01] e [Mat03]

$$h_{i,j+1} = \frac{1}{\alpha_j}\left(\frac{\mu_{i+1}}{\mu_i}\alpha_i h_{i+1,j} + (\beta_i - \beta_j)h_{i,j} + \frac{\mu_{i-1}}{\mu_i}\gamma_i h_{i-1,j} - \gamma_j h_{i,j-1}\right), \; i,j \geq 1 \tag{2.29}$$

onde $\mu_i = \|\varphi_i\|_w^2$, $i \geq 0$. A relação (2.29) possui a seguinte "estrutura",

$$\begin{array}{ccc} & h_{i-1,j} & \\ h_{i,j-1} & h_{i,j} & \mathbf{h_{i,j+1}} \\ & h_{i+1,j} & \end{array}$$

permitindo que o elemento assinalado a negrito seja calculado à custa dos outros quatro elementos. Para se inicializar a relação de recorrência recorre-se à definição (2.19). Se $\varphi_0(x) = k_0$ então,

$$h_{i,0} = k_0 f_i, \qquad i \geq 0, \tag{2.30}$$

$$h_{j,i} = \frac{\mu_i}{\mu_j}h_{i,j}, \qquad i,j \geq 0. \tag{2.31}$$

Para se determinar todos os coeficientes $h_{i,j}$ envolvidos em (2.26)-(2.27)

$$\left[\begin{array}{c|c} \mathbf{G_{p,q}} & \mathbf{g_{p,q}} \\ \hline \mathbf{H_{p,q}} & \mathbf{h_{p,q}} \end{array}\right] = \left[\begin{array}{ccc|c} h_{0,0} & \cdots & h_{0,q-1} & h_{0,q} \\ \vdots & & \vdots & \vdots \\ h_{p,0} & \cdots & h_{p,q-1} & h_{p,q} \\ \hline h_{p+1,0} & \cdots & h_{p+1,q-1} & h_{p+1,q} \\ \vdots & & \vdots & \vdots \\ h_{p+q,0} & \cdots & h_{p+q,q-1} & h_{p+q,q} \end{array}\right]$$

necessitamos de calcular os elementos $h_{i,j}$, $i = 0, \ldots, p + 2q - j$, $j = 0, \ldots, q$

$$\mathbf{HG}_{p,q} = \left[\begin{array}{ccc|c} h_{0,0} & h_{0,1} & \cdots & h_{0,q-1} & h_{0,q} \\ \vdots & \vdots & & \vdots & \\ h_{p,0} & h_{p,1} & \cdots & h_{p,q-1} & h_{p,q} \\ \hline h_{p+1,0} & h_{p+1,1} & \cdots & h_{p+1,q-1} & h_{p+1,q} \\ \vdots & \vdots & & \vdots & \\ h_{p+q,0} & h_{p+q,1} & \cdots & h_{p+q,q-1} & h_{p+q,q} \\ \hline\hline h_{p+q+1,0} & h_{p+q+1,1} & \cdots & h_{p+q+1,q-1} & \\ \vdots & \vdots & \iddots & & \\ h_{p+2q-1,0} & h_{p+2q-1,1} & & & \\ h_{p+2q,0} & & & & \end{array}\right]$$

Este facto mostra que enquanto na AP linear, e não linear, de uma série de potências do tipo (p, q) a construção de um aproximante apenas exige o conhecimento dos coeficientes da série até à ordem $p + q$, na AP de séries de polinómios ortogonais a construção de um aproximante linear exige o conhecimento dos coeficientes, $f_i = h_{i,0}$, até à ordem $p + 2q$.

Iremos, de seguida apresentar resultados que permitem calcular os aproximantes de Padé lineares utilizados ao longo deste trabalho.

1. Aproximantes de Padé de expansões de Chebyshev, a que chamaremos de aproximantes de Chebyshev-Padé (ACP).

2. Aproximantes de Padé de expansões de Legendre, também denominados por aproximantes de Legendre-Padé (ALP).

Cálculo de APC lineares no intervalo $[-1, 1]$ Os polinómios de Chebyshev normalizados com a condição $T_i(1) = 1, \quad i \geq 0$, satisfazem a relação de recorrência (2.28) com

$$\begin{cases} \alpha_i = \gamma_i = \frac{1}{2}, \quad \beta_i = 0, \quad i \geq 1 \\ \alpha_0 = 1, \quad \beta_0 = 0 \end{cases}$$

e $\mu_0 = \pi$, $\mu_i = \pi/2, \quad i \geq 1$. Então, usando a relação de recorrência (2.29), inicializada com as relações (2.30) e (2.31) tem-se:

$$\begin{cases} h_{i,0} = f_i, \ i \geq 0 \\ h_{0,j} = \frac{1}{2} f_j, \ j \geq 1 \\ h_{1,1} = h_{0,0} + \frac{1}{2} h_{2,0} \\ h_{i,1} = \frac{1}{2} h_{i-1,0} + \frac{1}{2} h_{i+1,0}, \quad i \geq 2 \\ h_{1,j} = 2h_{0,j-1} + h_{2,j-1} - h_{1,j-2}, \quad j \geq 2 \\ h_{i,j} = h_{i-1,j-1} + h_{i+1,j-1} - h_{i,j-2}, \quad i, j \geq 2 \end{cases}$$

Cálculo de APC lineares num intervalo $[\mathbf{a},\mathbf{b}]$: As entradas $h_{p,q}$ da matriz $\mathbf{HG_{p,q}}$ podem calcular-se do seguinte modo: considerando $\left\{T_k^{[a,b]}(x)\right\}_{k\geq 0}$ a família dos polinómios de Chebyshev ortogonais no intervalo $[a,b]$, $a>b$, normalizados pela condição $T_k^{[a,b]}(b)=1,\ k\geq 0$ então, tem-se que os polinómios

$$T_k^{[a,b]}(x)=T_k^{[-1,1]}\left(\frac{2}{b-a}x-\frac{a+b}{b-a}\right),\ k\geq 0$$

satisfazem a relação de recorrência

$$T_0^{[a,b]}(x)=1$$

$$T_1^{[a,b]}(x)=\tfrac{2}{b-a}x-\tfrac{a+b}{b-a}$$

$$T_{k+1}^{[a,b]}(x)=2T_1^{[a,b]}(x)T_k^{[a,b]}(x)-T_{k-1}^{[a,b]}(x),\ k\geq 1$$

logo temos

$$xT_k^{[a,b]}(x)=\frac{b-a}{4}T_{k+1}^{[a,b]}(x)+\frac{a+b}{2}T_k^{[a,b]}(x)+\frac{b-a}{4}T_{k-1}^{[a,b]}(x),\ k\geq 1$$

e concluímos que: $\alpha_0=\frac{b-a}{2}$, $\alpha_k=\gamma_k=\frac{b-a}{4}, k\geq 1$ e $\beta_k=\frac{a+b}{2}, k\geq 0$. Mais, os quocientes $\frac{\mu_{i+1}}{\mu_i}$, $\frac{\mu_{i-1}}{\mu_i}$ que ocorrem na relação de recorrência (2.29) não dependem do intervalo $[a,b]$ dado que,

$$\begin{aligned}
&\int_a^b T_m^{[a,b]}(x)T_n^{[a,b]}(x)\frac{dx}{(1-u^2)^{1/2}},\quad \text{onde}\ \ u=\frac{2}{b-a}x-\frac{a+b}{b-a}\\
&=\frac{2}{b-a}\int_{-1}^{1}T_m^{[a,b]}\left(\frac{b-a}{2}u+\frac{a+b}{2}\right)T_n^{[a,b]}\left(\frac{b-a}{2}u+\frac{a+b}{2}\right)\frac{du}{(1-u^2)^{1/2}}\\
&=\frac{2}{b-a}\int_{-1}^{1}T_m^{[-1,1]}(u)T_n^{[-1,1]}(u)\frac{du}{(1-u^2)^{1/2}}=\begin{cases}0 & \text{se}\ \ n\neq m\\ \frac{2\pi}{b-a} & \text{se}\ \ n=m=0\\ \frac{\pi}{b-a} & \text{se}\ \ n=m\neq 0\end{cases}
\end{aligned}$$

ou seja, tem-se $\frac{\mu_1}{\mu_0}=2$ e $\frac{\mu_{i+1}}{\mu_i}=1$, para $i>0$.

Cálculo de ALP lineares Os polinómios de Legendre, normalizados com a condição $P_i(1)=1,\ \ i\geq 0$, satisfazem a relação de recorrência (2.28) com

$$\begin{cases}\alpha_i=\frac{i+1}{2i+1}, \quad \beta_i=0, \quad \gamma_i=\frac{i}{2i+3}, i\geq 1\\ \alpha_0=1, \quad \beta_0=0\end{cases}$$

e $\mu_i=\frac{2}{2i+1},\ \ i\geq 0$. Então, tem-se

$$\begin{cases} h_{i,0} = f_i, \ i \geq 0 \\ h_{0,j} = \frac{1}{2j+1} f_j, \ j \geq 1 \\ h_{i,1} = \frac{i+1}{2i+3} f_{i+1} + \frac{i}{2i-1} f_{i-1}, \ i \geq 1 \\ h_{1,j} = \frac{3}{2j+1} h_{j,1}, \ j \geq 2 \\ h_{i,j} = \frac{2j+1}{j+1} \left[\frac{i+1}{2i+3} h_{i+1,j} + \frac{i}{2i-1} h_{i-1,j} \right] - \frac{j}{j+1} h_{i,j-1}, \ i \geq 2, \ j \geq 1. \end{cases}$$

2.6.3 Cálculo de AP não lineares

Como foi observado anteriormente o cálculo dos coeficientes de AP não lineares exige geralmente a resolução de um sistema de equações algébricas não lineares. Estas equações podem ser estabelecidas da seguinte forma (ver por exemplo [Sid03]). Supondo, por uma questão de simplicidade, que os pólos de um AP não linear $R_{p,q}$, com $p \geq q-1$, têm todos multiplicidade um então decompondo $R_{p,q}$ em frações simples, ou seja,

$$R_{p,q}(z) = r(z) + \sum_{k=1}^{q} \frac{A_k}{z - \xi_k}, \quad r(z) = \sum_{k=0}^{p-q} r_k \varphi_k(z),$$

as condições (2.20) tomam a forma

$$r_k + \|\varphi_k\|_w^{-2} \sum_{i=1}^{q} A_i \psi_k(\xi_i) = f_k, \quad k = 0, 1, \ldots, p-q \tag{2.32}$$

$$\|\varphi_k\|_w^{-2} \sum_{i=1}^{q} A_i \psi_k(\xi_i) = f_k, \quad k = p-q+1, \ldots, p+q \tag{2.33}$$

onde as funções $\psi_k(\xi)$, $k = 0, 1. \ldots$ são as funções de segunda espécie definidas por

$$\psi_k(\xi) = \int_I \frac{\varphi_k(x)}{x - \xi} w(x) dx.$$

Note-se que impondo a condição $\xi_i \notin I$, $i = 1, \ldots, q$ fica assegurada a existência do AP não linear $R_{p,q}$. Caso contrário não existe AP não linear do tipo (p, q). Logo, caso exista $R_{p,q}$, podemos determinar os coeficientes A_i e ξ_i resolvendo as equações não lineares (2.32) e (2.33). O seguinte resultado [Sid03] garante a unicidade do AP não linear.

Teorema 2.6.2. Seja $f \in L_w^2(I)$ uma função real tal que $f(z) = \sum_{k=0}^{\infty} f_k \varphi_k(z)$, $f_k = \mathcal{U}_k(f)$, e seja $R_{p,q}$ um AP não linear do tipo (p, q) de f sem pólos no intervalo I. Então $R_{p,q}$ é único.

Em [Fle73] é dado um algoritmo que permite calcular aproximantes de Legendre-Padé não lineares. Em [CL74] é dado um algoritmo, que descrevemos de seguida, para calcular os chamados *aproximantes de Clenshaw-Lord* que, para valores de $p \geq q-1$, são equivalentes aos aproximantes de Chebyshev-Padé não lineares. Salientamos que este algoritmo, chamado de *algoritmo de Clenshaw-Lord*, apenas exige a resolução de um sistema de equações lineares.

Algoritmo de Clenshaw-Lord: Seja f uma função com expansão de Chebyshev $f(z) = {\sum_{k=0}^{\infty}}' f_k T_k(z)$, onde a *plica* no somatório indica que ${\sum_{k=0}^{n}}' a_k = \frac{a_0}{2} + \sum_{k=1}^{n} a_k$. E seja,

$$R_{p,q}(z) = \frac{{\sum_{k=0}^{p}}' a_k T_k(z)}{{\sum_{k=0}^{q}}' b_k T_k(z)},$$

o ACP não linear normalizado com a condição $b_0 = 2$. Então, os coeficientes a_k e b_k podem determinar-se do seguinte modo:

1. Resolver o sistema de equações lineares em ordem a γ_ℓ, $\ell = 1, \ldots, p$

$$\sum_{\ell=0}^{p} \gamma_\ell f_{|r-\ell|} = 0, \quad r = q+1, \ldots, q+p \ \text{ onde, } \gamma_0 = 1.$$

2. Calcular

$$b_\ell = \mu \sum_{i=0}^{q-\ell} \gamma_i \gamma_{\ell+i}, \quad \ell = 1, \ldots, q$$

   onde, $\mu^{-1} = \frac{1}{2}\sum_{i=0}^{q} \gamma_i^2$.

3. Calcular

$$a_r = \frac{1}{2} {\sum_{\ell=0}^{q}}' b_\ell \left(f_{r+\ell} + f_{|r-\ell|} \right), \quad r = 0, 1, \ldots, p.$$

2.7 Localização de pólos e zeros de AP de séries ortogonais perturbadas

Nesta secção generalizam-se os resultados obtidos na secção 2.5, relativa aos AP de séries de potências perturbadas. Iremos efetuar testes numéricos com séries de Chebyshev e de Legendre e compará-los com os resultados obtidos com séries de potências [MMR14]. Consideramos os ruídos (2.16) e (2.17) para séries de potências e, os ruídos análogos

$$T_{\varphi_\epsilon}(z) = \sum_{k=0}^{\infty} \epsilon r_k \varphi_k(z), \ \text{ do tipo I,} \tag{2.34}$$

$$T_{\varphi_\omega}(z) = \sum_{k=0}^{\infty} \omega \frac{r_k}{2^k} \varphi_k(z), \ \text{ do tipo II,} \tag{2.35}$$

para séries de polinómios ortogonais, onde os coeficientes aleatórios r_k satisfazem as mesmas condições que em (2.16) e (2.17). Iremos apenas analisar os casos particulares relativos aos polinómios de Chebyshev T_k e aos polinómios de Legendre P_k.

2.7.1 Localização de pólos e zeros de ACP

Tendo em vista analisarmos o comportamento dos zeros e pólos de ACP diagonais de expansões de Chebyshev considerámos duas funções $f(z) = 1/(z-2)$ e $g(z) = \ln(5/4-z)$ com singularidades análogas aos exemplos 2.5.1 e 2.5.2 respetivamente.

Exemplo 2.7.1. A função f possui série de Taylor centrada em $z = 0$, $S_f(z) = \sum_{k=0}^{\infty} c_k z^k$, $c_k = -2^{-(k+1)}$, e expansão de Chebyshev $C_f(z) = \sum_{k=0}^{\infty} a_k T_k(z)$, onde os coeficientes de Chebyshev a_k são determinados usando o seguinte resultado [D.64],

Proposição 2.7.1. Seja f uma função racional com M pólos simples, $z_\ell \notin [-1, 1]$, com resíduos $\mathrm{res}(z_\ell)$, $\ell = 1, \ldots, M$. Então, os coeficientes a_k da expansão de Chebyshev da função f são determinados por,

$$a_k = -2 \sum_{\ell=1}^{M} \frac{\mathrm{res}(z_\ell)}{\sqrt{z_\ell^2 - 1}\left(z_\ell \pm \sqrt{z_\ell^2 - 1}\right)^k}, \tag{2.36}$$

onde a escolha do sinal, em cada parcela, é feita de modo a que se verifique a seguinte desigualdade $\left|z_\ell \pm \sqrt{1 - z_\ell^2}\right| > 1$.

Generalizou-se o procedimento efetuado por Froissart, ver secção 2.5, introduzindo ruídos do tipo I e do tipo II às séries S_f e C_f. Determinamos os pólos e zeros de aproximantes de Taylor-Padé[1] (ATP) diagonais da série de potências perturbada e determinamos os pólos e zeros dos ACP lineares diagonais e dos ACP não lineares diagonais. Os resultados obtidos para a localização dos pólos e zeros dos ATP são análogos aos obtidos no exemplo 2.5.1 para a localização dos pólos e zeros dos ATP da função racional $1/(1-z)$ (com a exceção de que o pólo situa-se, neste exemplo, no ponto $z = 2$).

Os pólos e zeros dos ACP lineares e dos ACP não lineares exibem os seguintes padrões:

Ruído do tipo I:

(1) Existe um pólo estável, ξ_1, perto do pólo da função f, $z = 2$, tal que $|\xi_1 - 2| = \mathcal{O}(\epsilon)$ para os aproximantes $\Phi_{1,1}$ e $R_{1,1}$ e $|\xi_1 - 2|$ diminui quando se aumenta a ordem dos ACP.

[1] Sempre que possam existir eventuais confusões entre os aproximantes de Padé de séries de potências e os aproximantes de Padé de séries ortogonais, chamaremos aos primeiros *aproximantes de Taylor-Padé* (ou ATP).

(2) Existe um zero instável (zero fantasma), η_1, tal que $|\eta_1| = \mathcal{O}(\epsilon^{-1})$ para os aproximantes $\Phi_{1,1}$, $R_{1,1}$ e $|\eta_2|$ aumenta quando se aumenta a ordem dos ACP.

(3) Para $n > 1$ os pólos e zeros restantes, ξ_k e η_k, $k = 2, \ldots, n$, formam pares de Froissart artificiais, a distância entre ξ_k e η_k diminui quando se aumenta a ordem dos ACP e localizam-se perto do intervalo real $[-1, 1]$.

Ruído do tipo II**:**

(1) Analogamente existe um pólo estável e um zero fantasma que satisfazem as mesmas propriedades que o pólo estável e o zero fantasma do exemplo anterior.

(2) Para $n > 1$ os pólos e zeros restantes formam igualmente pares de Froissart mas localizam-se perto da elipse de Bernstein $\mathcal{E}_2 = \left\{z : \left|z + \sqrt{z^2 - 1}\right| = 2\right\}$.

Representamos estes resultados nas Figuras 2.3, para perturbações do tipo I, e 2.4, para perturbações do tipo II.

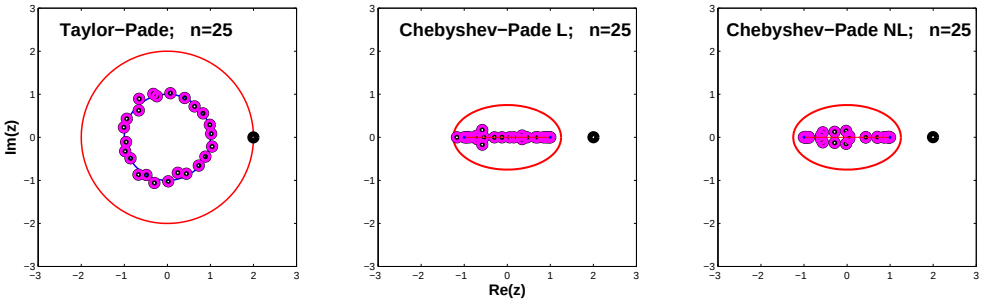


Figura 2.3: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de AP diagonais, com $n = 25$ da função f perturbada com ruído do tipo I e $\epsilon = 10^{-4}$. Da esquerda para a direita: ATP, ACP linear e ACP não linear. Para facilitar a leitura da figura não incluímos o zero fantasma.

Exemplo 2.7.2. Consideramos a função $g(z) = \ln(5/4 - z)$, a qual tem dois pontos de ramificação em $z = 5/4$ e em $z = \infty$ e uma linha de ramificação, que une os dois pontos

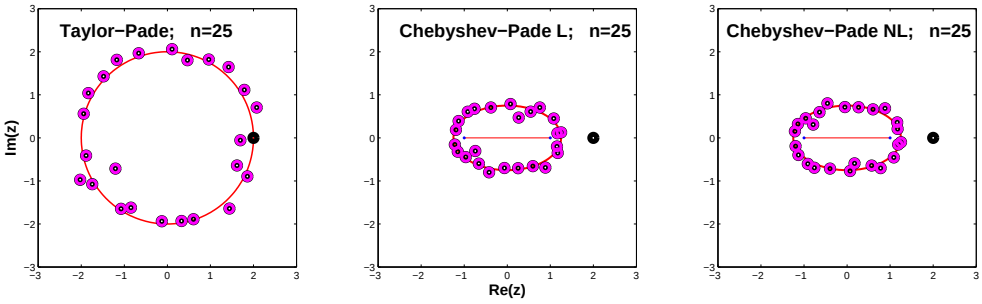


Figura 2.4: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de AP diagonais, com $n = 25$ da função f perturbada com ruído do tipo II e $\omega = 10^{-4}$. Da esquerda para a direita: ATP, ACP linear e ACP não linear. Para facilitar a leitura da figura não incluímos o zero fantasma.

de ramificação. A série de Taylor, centrada em $z = 0$, e a série de Chebyshev, [AS65] da função g são dadas respetivamente por,

$$S_g(z) = \ln(5/4) - \sum_{K=1}^{\infty} \frac{4^k}{5^k k} z^k, \quad \text{e} \quad C_g(z) = -\sum_{K=1}^{\infty} \frac{1}{2^k k} T_k(z).$$

Todos os AP calculados, ATP, ACP Lineares e ACP não lineares, usando as séries perturbadas com ruídos do tipo I e do tipo II, possuem zeros e pólos que mimetizam a estrutura das singularidades da função g. Ou seja: possuem um pólo estável perto do ponto de ramificação $z = 5/4$, possuem um zero estável perto do zero da função g, $z = 1/4$, e têm um conjunto de pólos e zeros que se intercalam e representam a linha de ramificação da função g. Os pólos e zeros restantes dos ATP formam pares de Froissart e localizam-se perto da fronteira natural do ruído, na circunferência de raio 1 para o ruído do tipo I e na circunferência de raio 2 para o ruído do tipo II. Os pólos e zeros restantes dos ACP (lineares e não lineares) formam igualmente pares de Froissart e localizam-se perto do segmento $[-1, 1]$ para o ruído do tipo I e da elipse de Bernstein $\mathcal{E}_2$ para o ruído do tipo II. Este comportamento é ilustrado nas Figuras 2.5 e 2.6.

Observação: *É interessante observar que existe uma relação entre a localização dos pares de Froissart dos ATP, que mimetizam a fronteira natural dos ruídos* (2.16) *e* (2.17)*, e os pares de Froissart dos ACP lineares e não*

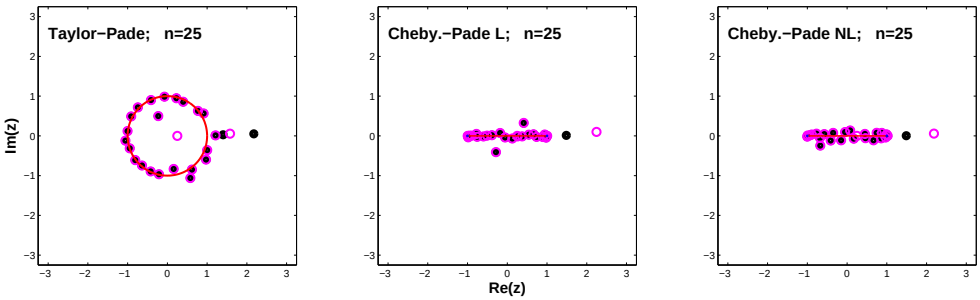


Figura 2.5: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de AP diagonais, $n = 25$ da função g perturbada com ruído do tipo I, $\epsilon = 10^{-4}$. Da esquerda para a direita: ATP, ACP linear e ACP não linear. Para facilitar a leitura da figura não incluímos todos os pólos e zeros que representam a linha de ramificação da função g.

> *lineares. Como a transformação de Joukowski, J, definida no plano complexo por $J(z) = \frac{z+z^{-1}}{2}$, envia circunferências centradas na origem de raio $\rho \geq 1$ em elipses de Bernstein $\mathcal{E}_\rho$ (a imagem da circunferência de raio unitário é a "elipse" de Bernstein degenerada $\mathcal{E}_1 = [-1, 1]$).*

Esta observação justifica as seguintes conjecturas.

Conjectura 2.1. As séries aleatórias de Chebyshev, (2.34) e (2.35), representam, formalmente, funções com fronteiras naturais no segmento $[-1, 1]$ e na elipse $\mathcal{E}_2$, respetivamente.

Conjectura 2.2. Os pares de Froissart de ACP, originados pelos ruídos (2.34) e (2.35), localizam-se na imagem da transformação de Joukowski das fronteiras naturais dos ruídos (2.16) e (2.17) respetivamente.

Foram efectuados testes idênticos com séries aleatórias de Legendre, (2.34) e (2.35). Os resultados obtidos foram análogos aos resultados obtidos com séries aleatórias de Chebyshev. Consequentemente generalizamos a conjectura 2.1 a séries aleatórias de Chebyshev e de Legendre, (2.34) e (2.35), e a conjectura 2.2 a ACP e ALP.

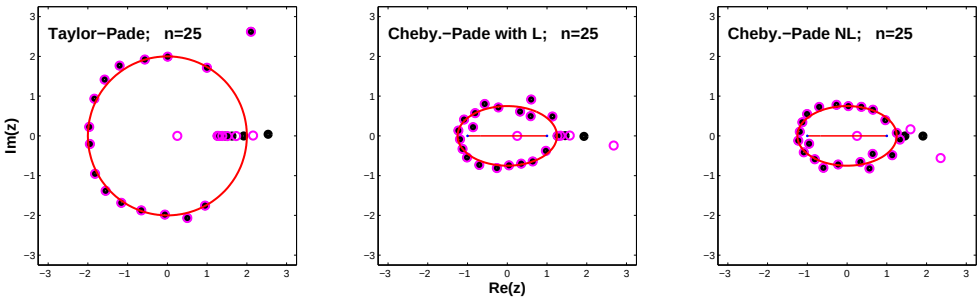


Figura 2.6: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de AP diagonais, com $n = 25$ da função g perturbada com ruído do tipo II e $\omega = 10^{-4}$. Da esquerda para a direita: ATP, ACP linear e ACP não linear. Para facilitar a leitura da figura não incluímos todos os pólos e zeros que representam a linha de ramificação da função g.

2.7.2 Localização de pólos e zeros de ALP

Consideramos as funções f e g definidas por, $f(z) = \frac{2}{\sqrt{5-4z}}$ e $g(z) = -\ln((1-z)/2)$ respetivamente. A série de Legendre da função f é determinada pela função geradora [AS65]

$$\frac{1}{\sqrt{1-2\alpha z+\alpha^2}} = \sum_{k=0}^{\infty} \alpha^k P_k(z) \tag{2.37}$$

fazendo $\alpha = 1/2$. Ou seja, $f(z) = \sum_{k=0}^{\infty} 1/2^k P_k(z)$. A função f tem dois pontos de ramificação em $z = 5/4$ e $z = \infty$ e uma linha de ramificação que une os pontos de ramificação. A função g tem um zero em $z = -1$, dois pontos de ramificação, em $z = 1$ e em $z = \infty$, e a sua série de Legendre é dada por [Hol69][2],

$$g(z) = 1 + \sum_{k=1}^{\infty} \frac{2k+1}{k(k+1)} P_k(z).$$

Os resultados são idênticos aos resultados obtidos com séries de Chebyshev. Ou seja, para a função f perturbada com um ruído do tipo I ou do tipo II os ALP diagonais possuem um pólo estável perto do ponto de ramificação $z = 5/4$, um zero instável cujo módulo cresce com a ordem do ALP e um conjunto de pólos e de zeros que se intercalam

[2]Esta referência contém um erro na equação (52), deve substituir-se $2k-1$ por $2k+1$.

e que representam a linha de ramificação de f. Além disso, os pólos e zeros restantes constituem pares de Froissart e localizam-se perto do segmento $[-1, 1]$ se a perturbação for do tipo I e localizam-se perto da elipse se a perturbação for do tipo II. Para a função g perturbada os ALP diagonais possuem um pólo estável perto do ponto de ramificação $z = 1$, um zero estável perto do zero de g, $z = -1$, um conjunto de zeros e pólos que se intercalam e que representam a linha de ramificação de g e os pólos restantes mimetizam os ruídos e possuem as mesmas propriedades observadas nos exemplos anteriores. Este comportamento é ilustrado nas Figuras 2.7 e 2.8

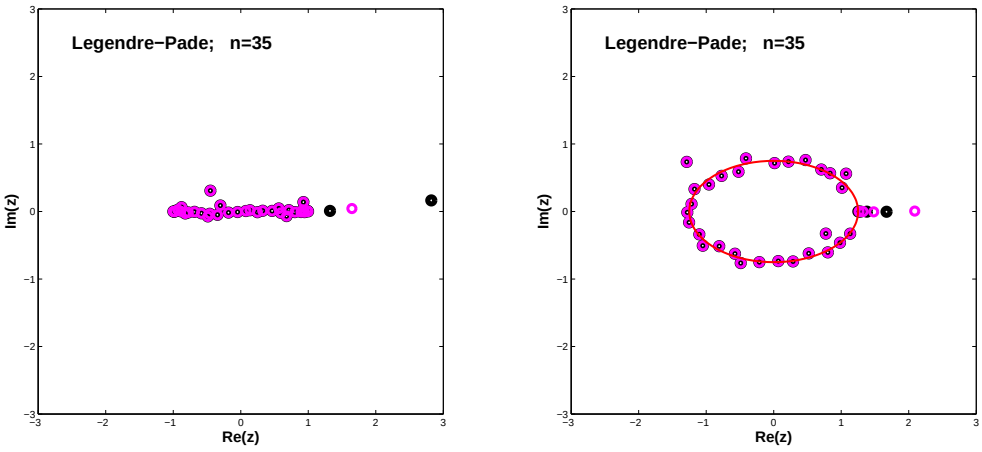


Figura 2.7: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de ALP diagonais, com $n = 35$ da função f perturbada com ruído do tipo I (à esquerda) e II (à direita), $\epsilon = \omega = 10^{-4}$. Para facilitar a leitura da Figura não incluímos todos os pólos e zeros que representam a linha de ramificação da função f.

2.8 Utilização dos pares de Froissart na deteção de um "bom" AP

O aparecimento de pares de Froissart é um indicador de que há instabilidade no cálculo dos aproximantes de Padé, [BM14]. Para ultrapassar este inconveniente Gonnet *et al*, [GGT13], introduziram o conceito de *aproximante de Padé robusto* para séries de potências. Basicamente um aproximante de Padé robusto não possui pólos espúrios nem pares de Froissart. Estas características dos aproximantes de Padé robustos estabilizam o seu cálculo, dado que a eliminação dos pares de Froissart torna as matrizes envolvidas bem condicionadas [BM14].

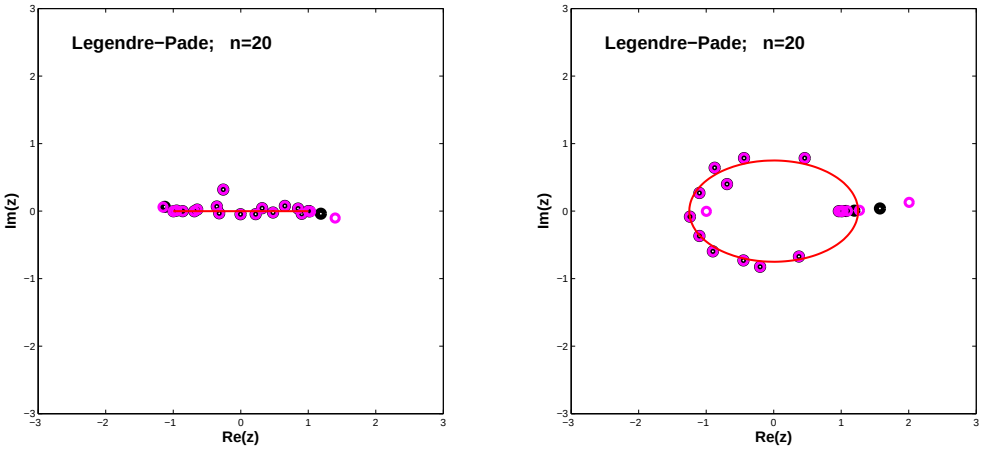


Figura 2.8: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de ALP diagonais, $n = 20$ da função g perturbada com ruído do tipo I (à esquerda) e II (à direita), $\epsilon = \omega = 10^{-3}$. Para facilitar a leitura da Figura não incluímos todos os pólos e zeros que representam a linha de ramificação da função g.

Neste trabalho, introduzimos uma estratégia para a escolha de um "bom" AP de séries ortogonais. Esta estratégia, baseia-se no conceito de aproximante de Padé robusto atrás mencionado, e consiste no seguinte procedimento:

1. fixamos $\text{tol} > 0$, uma quantidade "pequena" a que chamamos de *tolerância*;
2. para cada AP, $N_{p,q}/D_{p,q}$, calculamos os pólos ξ_k, $k = 1, \ldots, q$ e os zeros η_k, $k = 1, \ldots, p$;
3. contamos o número, $n_{p,q}$, de pares de Froissart (ξ_k, η_ℓ), pares de pólos e zeros, cuja distância entre si é menor do que a tolerância

$$n_{p,q} = \#\left\{(\xi_k, \eta_\ell) \ : \ N_{p,q}(\eta_\ell) = 0 \ \wedge \ D_{p,q}(\xi_k) = 0 \ \wedge \ |\xi_k - \eta_\ell| < \text{tol}\right\};$$

4. dispomos os valores $n_{p,q}$ numa tabela, que designamos por *tabela de Froissart*;
5. identificamos na tabela de Froissart, regiões onde não existem pares de Froissart, ou seja regiões onde $n_{p,q} = 0$;
6. um "bom" aproximante satisfaz as condições,

$$n_{p,q} = 0 \ \wedge \ p + q \text{ é máximo } \wedge \ |p - q| \text{ é mínimo } \wedge \ D_{p,q} \neq 0 \text{ em } [-1, 1]. \quad (2.38)$$

Observação: *A escolha do melhor AP, no sentido de ser o AP que minimiza uma dada norma do erro, na região livre de pares de Froissart não é fácil. Pode ocorrer que o melhor AP não verifique as condições 2.38. Contudo, nos testes efetuados, o nosso procedimento este procedimento revelou-se eficaz.*

O exemplo seguinte é paradigmático no procedimento adotado ao longo deste trabalho.

Exemplo 2.8.1. Consideramos a função f_α e a sua expansão de Legendre [GR07][3]

$$f_\alpha(x) = \frac{1-\alpha^2}{(1+\alpha^2-2\alpha x)^{3/2}} = 1 + \sum_{k=1}^{\infty}(2k+1)\alpha^k P_k(x), \quad |\alpha| < 1.$$

Para $\alpha = 1/2$ a função tem dois pontos de ramificação em $x = 5/4$ e em $x = \infty$ e uma linha de ramificação que os une. Pretende-se encontrar o ALP que melhor aproxima esta função. Para o efeito usamos os coeficientes da expansão de Legendre que consideramos exatos, na medida que são calculados com erro igual à precisão do software usado. Logo, nesta experiência, o aparecimento de pares de Froissart deve-se somente aos erros causados na resolução do sistema de equações lineares (2.26). Noutras palavras, os erros nos coeficientes dos polinómios do denominador dos ALP dependem fortemente do número de condição da matriz (2.25). Notamos, igualmente, que estes erros estão correlacionados com os erros nos coeficientes do denominador dado que, os coeficientes do numerador obtém-se dos coeficientes do denominador usando a relação (2.27). Obviamente estas observações são extensíveis a todos os AP lineares.

Na Figura 2.9 apresentamos a tabela de Froissart para valores de $n_{p,q}$, com $p, q = 1 \ldots 20$, com uma tolerância $tol = 10^{-3}$.
Para verificar a estabilidade da região livre de pares de Froissart construimos as tabelas de Froissart para os valores tol $= 10^{-2}$ e tol $= 10^{-3}$ e verificou-se que a região livre de pares de Froissart permanece inalterável, apenas são alterados alguns valores de $n_{p,q} > 0$. Note-se que a região livre de pares de Froissart está separada da região com pares de Froissart e que na região com pares de Froissart o número de pares de Froissart cresce, em geral, quando o valor de $p + q$ cresce. Este padrão verificou-se em quase todos os exemplos considerados neste trabalho.

O próximo passo, neste procedimento, é encontrar um bom aproximante de Padé na região livre de pólos da função f. Como foi dito atrás vamos escolher um aproximante $\Phi_{p,q}$ na região livre de pares de Froissart, tal que $p+q$ seja máximo e que $|p-q|$ seja mínimo. Considerando uma sequência diagonal de ALP verificamos que os ALP, $\Phi_{p,p}$, não possuem pares de Froissart para valores de $p \leq 9$ e possuem pares de Froissart para valores de p superiores a 9. Logo, $\Phi_{9,9}$ será, em princípio, um candidato a melhor aproximante da sequência livre de pares Froissart.

[3]Foi retificado, nesta referência, um erro nesta igualdade (p. 989)

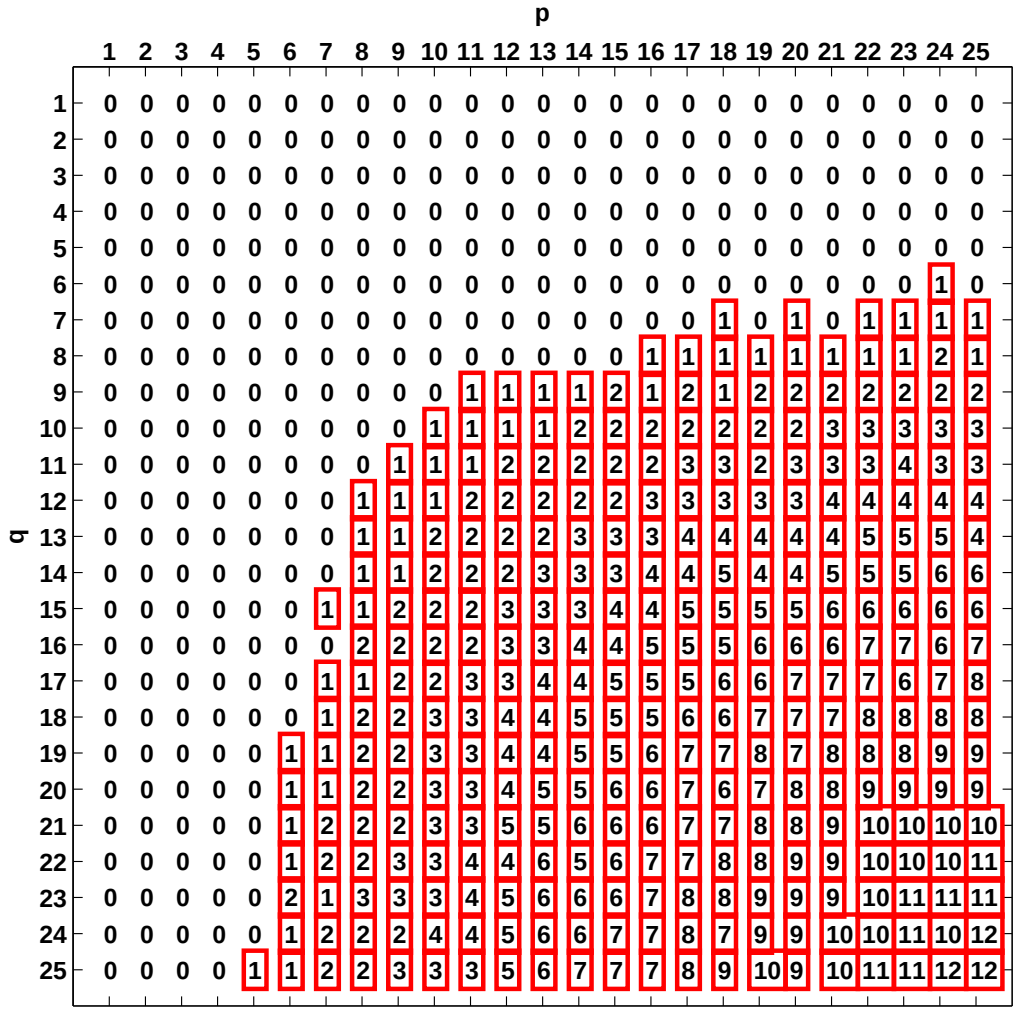


Figura 2.9: Tabela de Froissart com tolerância de 10^{-3} da função f_α, com $\alpha = 1/2$. As entradas com pares de Froissart estão assinaladas com um rectângulo de cor vermelha.

Note-se que não podemos afirmar categoricamente ser este um bom AP porque poderão existir um ou mais pólos no intervalo $[-1, 1]$ que não possuem zeros a uma distância inferior a 10^{-2}. Podemos verificar, para este exemplo, na Figura 2.11 que $\Phi_{9,9}$ é realmente um bom aproximante porque não possui pólos no intervalo $[-1, 1]$ (ver, Figura 2.10) e o erro absoluto, $\left|f_{1/2}(x) - \Phi_{9,9}(x)\right|$, é da ordem de 10^{-16} (da precisão da máquina) para valores de x próximos de $x = -1$ e crescem à medida que x se aproxima de $x = 1$ onde atinge o seu valor máximo, da ordem de 10^{-10}.

No próximo passo iremos usar os seguintes conceitos.

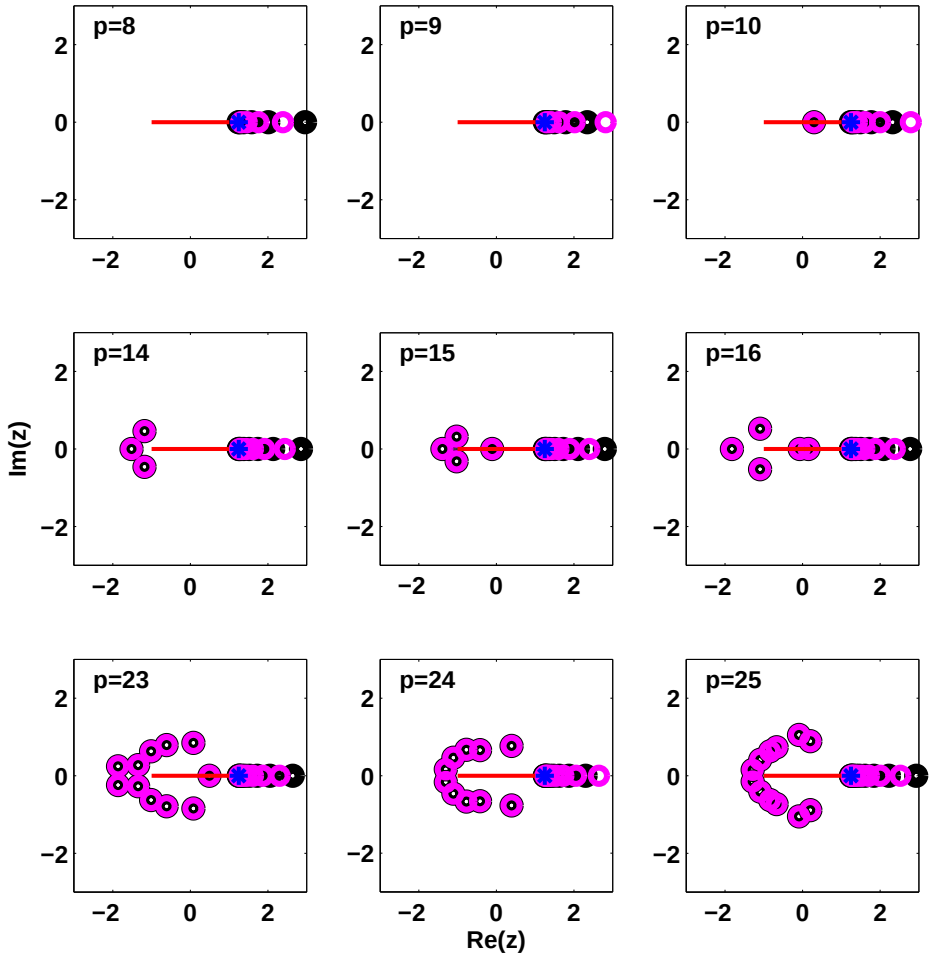


Figura 2.10: Localização do zeros (círculos a magenta) e dos pólos (pontos a preto) de ALP diagonais $\Phi_{p,p}$ da função $f_{1/2}$. O asterisco a azul representa o ponto de ramificação $z = 5/4$ e a linha vermelha o intervalo $[-1, 1]$. Para facilitar a leitura das imagens não incluímos todos os zeros e pólos que mimetizam a linha de ramificação de $f_{1/2}$.

Definição 2.8.1. Seja $[p/q]_f$ um AP de uma função f. Designamos, *grau numérico do numerador* de $[p/q]_f$ ao número

$$\nu^N_{p,q} = p - n_{p,q} \tag{2.39}$$

e designamos *grau numérico do denominador* de $[p/q]_f$ ao número

$$\nu^D_{p,q} = q - n_{p,q}, \tag{2.40}$$

no caso de um AP diagonal referimos apenas que o AP $[p/p]_f$ tem *grau numérico*

$$\nu_{p,p} = p - n_{p,q}. \tag{2.41}$$

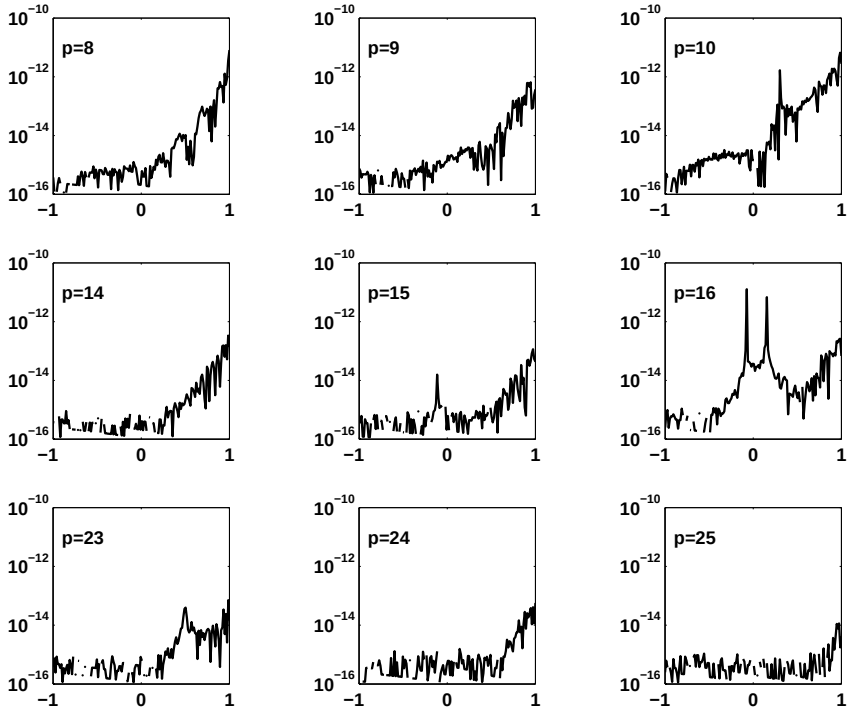


Figura 2.11: Erros absolutos de ALP diagonais $\Phi_{p,p}$ da função $f_{1/2}$.

Apesar de $\Phi_{9,9}$ já ser um bom aproximante da função $f_{1/2}$ podemos ainda, eventualmente, encontrar um aproximante na região da tabela de Froissart em que $n_{p,q} > 0$ que diminua o erro para valores de x próximos de $x = 1$. Com efeito, os valores dos graus numéricos $\nu_{p,p}$, $p \geq 0$, não crescem com p. Logo, se existir um aproximante $\Phi_{p,p}$, $p > 9$, sem pólos no intervalo $[-1, 1]$ é ainda possível melhorar a aproximação dada por $\Phi_{9,9}$. Pode-se ainda observar, que para valores de $p \in \mathcal{F} = \{10, 15, 16, 23\}$ os ALP $\Phi_{p,p}$ têm pares de Froissart no intervalo $[-1, 1]$. Deste modo deve-se escolher um $\Phi_{p,p}$ tal que $p \notin \mathcal{F}$ e que possua grau numérico $\nu_{p,p}$ máximo, de forma a melhorar a aproximação dada por $\Phi_{9,9}$. Podemos observar este comportamento analisando as Figuras 2.10 e 2.11. De facto a sequência de ALP diagonais tais que $p \notin \mathcal{F}$ vai melhorando a aproximação à medida que o valor de p aumenta. Podemos observar que para o último valor de p disponível o erro absoluto $\left|f_{1/2}(x) - \Phi_{25,25}(x)\right|$ é da ordem de 10^{-13}, para valores de x perto de um.

2.9 Estimação de Singularidades via AP de séries ortogonais

Nesta secção abordamos as versões para expansões de polinómios ortogonais obtidas por Buslaev em 2006, ver [Bus06], dos resultados obtidos por Fabry e Suetin para séries de potências, ver subsecção 2.4. Apenas incluímos os resultados relativos à segunda coluna da tabela de Padé de AP de expansões de polinómios ortogonais. Os resultados relativos a todas as sucessões coluna foram estabelecidos em 2009, pelo mesmo autor, ver [Bus09]

2.9.1 O problema inverso para séries ortogonais

Seja $\sigma(x)$ uma medida positiva, limitada, monótona crescente e com um número infinito de pontos de crescimento no intervalo $[-1,1]$ e cuja derivada $\sigma'(x)$ é não negativa e o conjunto $\{x \in [-1,1] \mid \sigma'(x) = 0\}$ tem medida nula, no sentido de Lebesgue.

O teorema 2.4.1, possui uma versão análoga relativa a séries ortogonais [Bus06]

Teorema 2.9.1. Seja $\{\phi_k(z)\}_{k\geq 0}$ uma família de polinómios ortonormados no intervalo $[-1,1]$ relativamente à medida σ onde σ satisfaz a condição de Szegö

$$\int_{-1}^{1} \frac{\ln \sigma'(x)}{\sqrt{1-x^2}} dx > -\infty. \tag{2.42}$$

Dada uma série ortonormal $\sum_{k\geq 0} c_k \phi_k(z)$ cujos coeficientes são tais que $\lim_{n\to\infty} c_n/c_{n+1}$ existe e é igual a λ, $|\lambda| > 1$. Então, a série converge uniformemente no conjunto

$$E_{|\lambda|} = [-1,1] \cup \left\{ z \in \mathbb{C} \mid \left| z + \sqrt{z^2-1} \right| < |\lambda| \right\}$$

e $\dfrac{\lambda^2+1}{2\lambda}$ é um ponto singular da função $f(z) = \sum_{k\geq 0} c_k \phi_k(z)$ que se encontra na fronteira de $E_{|\lambda|}$.

O seguinte resultado é a versão do teorema 2.4.2 para séries de polinómios ortonormais.

Teorema 2.9.2. Seja $\{\phi_k(z)\}_{k\geq 0}$ uma família de polinómios ortonormados no intervalo $[-1,1]$ relativamente à medida σ , onde σ satisfaz a condição de Szegö (2.42). Dada uma série $\sum_{k\geq 0} c_k \phi_k(z)$ tal que os pólos dos AFP da segunda coluna da tabela de Padé, $x_{p,1}$ tendem para η, quando $p \to \infty$ então se:

1. $\eta \in [-1,1]$, tem-se $\overline{\lim_{n\to\infty}} |c_n|^{1/n} = 1$;
2. $\eta \in \mathbb{C}\setminus[-1,1]$, tem-se que o limite $\lim_{n\to\infty} c_{n-1}/c_n$ existe e é igual a um destes números $\eta \pm \sqrt{\eta^2-1}$.

Observação: *Se acrescentarmos às condições do teorema (2.9.2) a condição de que a série* $\sum_{k\geq 0} c_k\phi_k(x)$ *não é formal numa vizinhança de* $[-1,1]$ *então* $\overline{\lim_{n\to\infty}}\,|c_n|^{1/n} < 1$ *e como* $|\eta - \sqrt{\eta^2-1}| < 1$ *ter-se-á necessariamente*

$$\lim_{n\to\infty} c_{n-1}/c_n = \eta + \sqrt{\eta^2-1}.$$

Logo, a série $\sum_{k\geq 0} c_k\phi_k(x)$ *converge uniformemente no interior de uma elipse com focos* ± 1 *e o ponto* η *é um ponto singular da função* $f(z) = \sum_{k\geq 0} c_k\phi_k(z)$.

Teorema 2.9.3 (Buslaev, [Bus06])**.** Seja f uma função holomorfa numa vizinhança $\mathcal{V}$ do intervalo $I = [-1,1]$ e $\{\phi_n(z)\}$ uma família de polinómios ortonormados no intervalo I relativamente à medida σ , onde σ satisfaz a condição de Szegö (2.42). Seja ainda, $\sum_{k\geq 0} c_k\phi_k(z)$ a série ortogonal que representa a função f em $\mathcal{V}$ e a função $g(z) = f\left(\frac{z+z^{-1}}{2}\right) = \sum_{k\geq 0} d_k\phi_k(z)$. Então, são equivalentes as seguintes proposições:

1. existe o limite $\lim_{n\to\infty} c_n/c_{n+1} = \lambda$.
2. existe o limite $\lim_{p\to\infty} x_{p,1} = \eta$, onde $x_{p,1}$ são os pólos dos AFP da forma $[p/1]_f$.
3. existe o limite $\lim_{n\to\infty} d_n/d_{n+1} = \lambda$, e tem-se: $\eta \in \mathbb{C}\setminus I$, $|\lambda| > 1$ e $\lambda = \eta + \sqrt{\eta^2-1}$.

Cálculo dos pólos dos AP da primeira coluna

Seja $\{\phi_k\}_{k\geq 0}$ é uma família de polinómios ortonormados. Por definição um AP linear $[p/q]_f$ de uma função $f(z) = \sum_{k\geq 0} c_k\phi_k(z)$ é uma função racional $N_{p,q}/D_{p,q}$, onde: $\mathrm{gr}\,(N_{p,q}) \leq p$, $\mathrm{gr}\,(D_{p,q}) \leq q$, $D_{p,q} \not\equiv 0$ e

$$\int_{-1}^{1} \left(D_{p,q}(x)f(x) - N_{p,q}(x)\right)\phi_k(x)d\sigma(x) = 0, \quad k = 0,1,\ldots,p+q.$$

Deste modo, os pólos, $x_{p,1}$, dos AP lineares da forma $[p/1]_f$ são dados por

$$x_{p,1} = \frac{\int_{-1}^{1} xf(x)\phi_{p+1}(x)d\sigma(x)}{\int_{-1}^{1} f(x)\phi_{p+1}(x)d\sigma(x)}$$

e supondo que os polinómios $\{\phi_k(z)\}_{k\geq 0}$ satisfazem a relação de recorrência

$$z\phi_n(z) = \alpha_n\phi_{n+1}(z) + \beta_n\phi_n(z) + \gamma_n\phi_{n-1}(z) \tag{2.43}$$

tem-se

$$x_{p,1} = \frac{\alpha_{p+1}c_{p+2} + \beta_{p+1}c_{p+1} + \gamma_{p+1}c_p}{c_{p+1}}. \tag{2.44}$$

Neste trabalho estamos interessados em estimar singularidades, usando um conjunto finito de coeficientes aproximados de expansões de polinómios ortogonais. Se a família de

polinómios, $\{\phi_k\}_{k\geq 0}$, for ortogonal e se considerarmos $\hat{f}(x) = \sum_{k=0}^{N} \hat{c}_k \phi_k(x)$ uma aproximação de uma função $f(x) = \sum_{k\geq 0} c_k \phi_k(x)$. Então a relação (2.44) toma a forma

$$x_{p,1} = \frac{\alpha_{p+1}\mu_{p+2}\hat{c}_{p+2} + \beta_{p+1}\mu_{p+1}\hat{c}_{p+1} + \gamma_{p+1}\mu_p\hat{c}_p}{\mu_{p+1}\hat{c}_{p+1}} \tag{2.45}$$

onde $\mu_k = s_k \int_{-1}^{1} \phi_k^2 d\sigma$, com $s_0 = 1/2$ e $s_k = 1$ para todo $k \geq 1$. Geralmente, para valores de N suficientemente grandes, é de esperar que os coeficientes $\hat{c}_k$ estejam suficientemente próximos dos coeficientes c_k, $k = 0, 1, \ldots, N$. Será, deste modo, expectável que o pólo do AP $[p/1]_{\hat{f}}$, seja uma boa aproximação do pólo de $[p/1]_f$ e consequentemente uma boa aproximação da singularidade, da função f, mais próxima do intervalo $[-1, 1]$.

Na próxima subsecção iremos encontrar fórmulas para calcular pólos de AP de expansões de Chebyshev de ACP da segunda e terceira coluna da tabela de Padé.

2.9.2 Estimativa de singularidades de expansões de Chebyshev

A família dos polinómios de Chebyshev satisfaz a relação de recorrência (2.43), com $\alpha_n = \gamma_n = 1/2$ e $\beta_n = 0$, $n \geq 1$. Logo a igualdade (2.45) toma a forma

$$x_{p,1} = \frac{c_{p+2} + c_p}{2c_{p+1}}, p \geq 1 \tag{2.46}$$

logo, se $c_{p+1} \neq 0$ podemos usar (2.46) para obter uma primeira estimativa da singularidade, de uma função, mais próxima do intervalo $[-1, 1]$.

No caso da função ter simetria (par ou ímpar) ou da função ter singularidades complexas será conveniente usar os pólos dos aproximantes de ACP da terceira coluna da tabela de Padé, por outra palavras, os dois pólos, $x_{p,2}^+$ e $x_{p,2}^-$ dos ACP $[p/2]$, $p = 0, 1, \ldots$. Estes pólos podem determinar-se usando a seguinte

Proposição 2.9.4. Seja f uma função representada pela expansão de Chebyshev $f(z) = \sum_{k=0}^{\infty} c_k T_k(z)$ e seja

$$\Delta_p = \begin{vmatrix} c_{p+1} & c_p + c_{p+2} \\ c_{p+2} & c_{p+1} + c_{p+3} \end{vmatrix} \neq 0.$$

Então os pólos $x_{p,2}^+$ e $x_{p,2}^-$ do ACP

$$[p/2]_{f(z)} = \frac{\sum_{k=0}^{p} a_k T_k(z)}{b_0 T_0(z) + b_1 T_1(z) + b_2 T_2(z)},$$

são dados por

$$x_{p,2}^{\pm} = \frac{-B_1 \pm \sqrt{B_1^2 - 8(B_0 - 1)}}{4} \tag{2.47}$$

com a convenção de que se obtém o pólo $x_{p,2}^+$ ($x_{p,2}^-$) se optarmos em (2.47) pela adição (subtração) , respetivamente. E onde

$$B_0 = -\frac{\begin{vmatrix} c_{p-1}+c_{p+3} & c_p+c_{p+2} \\ c_p+c_{p+4} & c_{p+1}+c_{p+3} \end{vmatrix}}{2\Delta_p} \quad \text{e} \quad B_1 = -\frac{\begin{vmatrix} c_{p+1} & c_{p-1}+c_{p+3} \\ c_{p+2} & c_p+c_{p+4} \end{vmatrix}}{\Delta_p}.$$

Demonstração: *Usando, as equações (2.24), a lei de multiplicação dos polinómios de Chebyshev* $T_m(x)T_n(x) = 1/2\left(T_{m+n} + T_{|m-n|}\right)$, $m, n \geq 0$ *e a normalização* $b_2 = 1$ *obtemos o sistema de duas equações lineares nas incógnitas* b_0 *e* b_1

$$\begin{cases} c_{p+1}b_0 + 1/2(c_p + c_{p+2})b_1 = -1/2(c_{p-1} + c_{p+3}) \\ c_{p+2}b_0 + 1/2(c_{p+1} + c_{p+3})b_1 = -1/2(c_p + c_{p+4}). \end{cases} \tag{2.48}$$

que é possível e determinado sse $\Delta_p \neq 0$. *Os pólos* $x_{p,2}^{\pm}$ *do ACP* $[p/2]_{f(z)}$ *são as raízes do polinómio* $D_{p,2}(z) = (b_0 - 1) + b_1 z + 2z^2$. *Resolvendo o sistema (2.48) tem-se que as raízes do polinómio* $D_{p,2}$ *são dadas pela relação (2.47).* ❑

2.9.3 Estimativa de singularidades de expansões de Legendre

A estimativa da singularidade de uma função f representada por uma expansão de Legendre pode ser efetuada usando os pólos $x_{p,1}$ de ALP $\Phi_{p,1}$, $p = 0, 1, \ldots$ usando a seguinte

Proposição 2.9.5. Seja f uma função representada por uma expansão de Legendre $f(x) \sim \sum_{k=0}^{\infty} c_k P_k(x)$ então, se $c_{p+1} \neq 0$, o pólo $x_{p,1}$ do ALP $\Phi_{p,1}$ é dado por

$$x_{p,1} = \frac{\frac{p+2}{2p+5}c_{p+2} + \frac{p+1}{2p+1}c_p}{c_{p+1}}. \tag{2.49}$$

Demonstração: *Usando a relação (A.15), verifica-se que a família dos polinómios de Legendre satisfaz a relação de recorrência (2.43), com* $\alpha_n = \frac{n+1}{2n+1}$, $\gamma_n = \frac{n}{2n+1}$ *e* $\beta_n = 0$. *Logo, como* $c_{p+1} \neq 0$, *a igualdade (2.45) toma a forma*

$$x_{p,1} = \frac{(p+2)\mu_{p+2}c_{p+2} + (p+1)\mu_p c_p}{(2p+3)\mu_{p+1}c_{p+1}}, \quad p = 1, 2, \ldots. \tag{2.50}$$

e usando a identidade $\mu_k = \|P_k\|^2 = \frac{2}{2k+1}$, *k=0,1,... em (2.50) obtém-se (2.49).*

❑

Exemplo 2.9.1. A função

$$f(x) = (\alpha^2 + 1 - 2\alpha x)^{-1/2}, \quad \alpha \in]-1, 1[$$

possui a expansão de Legendre, ver relação (2.37),

$$f(x) \sim 1 + \sum_{k=1}^{\infty} \alpha^k P_k(x), \quad x \in [-1, 1]$$

e possui a singularidade, para $\alpha \neq 0$, mais próxima da origem no ponto $\zeta = \frac{\alpha^2+1}{2\alpha}$. A relação (2.49) toma, neste caso, a forma

$$x_{p,1} = \alpha^{p-1}\left(\frac{p+2}{2p+5}\alpha^2 + \frac{p+1}{2p+1}\right) \tag{2.51}$$

e tem-se $\lim_{p\to\infty} x_{p,1} = \zeta$

Capítulo 3

Filtragem de Métodos Espectrais

Neste capítulo abordamos os problemas relacionados com o objetivo principal deste trabalho. Mais exatamente, descrevemos o método de *filtragem*[1] de soluções espectrais de equações diferenciais ordinárias usando para o efeito aproximações de Padé.

3.1 Introdução

Nesta secção iremos efetuar algumas observações sobre o método de filtragem e introduzir a notação usada nos exemplos descritos neste capítulo.

Podemos esquematizar este método de filtragem da seguinte forma,

Observação: *Existem outros métodos para filtrar soluções espectrais [Pey02]. Estes métodos baseiam-se, fundamentalmente, em efetuar mudanças de variáveis ou em decompor o domínio no qual pretendemos encontrar uma aproximação da solução de um dado problema, de forma a evitar singularidades ou evitar sub domínios nos quais a solução tenha variações bruscas. Não é nosso objetivo, neste trabalho, efetuar comparações entre estes métodos e o "nosso" procedimento. No entanto, como iremos verificar nos exemplos apresentados neste capítulo, salientamos que o uso de aproximantes de Padé, no processo de filtragem além de melhorar as aproximações dadas por soluções espectrais*

[1] Ao longo deste texto adotamos este termo, *filtragem*, como sendo a tradução da palavra inglesa *filtering*, que é usada, por vários autores, com o significado de melhorar a aproximação fornecida por um método espectral [Pey02].

possui ainda outras valências: localizar eventuais singularidades e a extensão analítica da solução espectral.

Os métodos espectrais, aplicados a problemas com soluções *suaves*, possuem taxa de convergência exponencial [CHQZ07]. Nestes casos não se justifica o uso de filtros dado que os erros cometidos pelas aproximações espectrais rapidamente atingem valores da ordem da precisão da máquina. Consequentemente, nesta secção, iremos apresentar apenas exemplos onde se justifique o uso de filtros. Ou seja, apresentamos problemas com soluções com singularidades no interior ou próximas do intervalo de ortogonalidade.

3.2 Erros cometidos no processo de filtragem.

Por ser extensa, começamos por estabelecer a notação a utilizar neste capítulo. No que diz respeito às funções, às aproximações polinomiais e respetivos erros designamos por:

$$
\begin{aligned}
&y(x) \sim \sum_{k=0}^{\infty} c_k \phi_k(x), \quad \text{a solução exata da equação diferencial na base } \{\phi_k\}_{k\geq 0}\,;\\
&y_N(x) = \sum_{k=0}^{N} c_k^{(N)} \phi_k(x), \quad \text{a solução espectral na base dos polinómios ortogonais;}\\
&\delta y_N(x) = y(x) - y_N(x), \quad \text{o erro da solução espectral;}\\
&\Delta y_N(x) = |\delta y_N(x)|\,, \quad \text{o erro absoluto da solução espectral;}\\
&\|\delta y_N\|_w = \left[(\delta y_N, \delta y_N)_w\right]^{1/2}, \quad \text{o erro na norma pesada da solução espectral;}\\
&\delta c_k^{(N)} = c_k - c_k^{(N)},\ k = 0, 1, \ldots, N, \quad \text{os erros dos coeficientes da solução espectral;}\\
&\Delta c_k^{(N)} = \left|\delta c_k^{(N)}\right|,\ k = 0, 1, \ldots, N, \quad \text{os erros absoluto dos coeficientes da solução espectral.}
\end{aligned}
$$

Relativamente às aproximações racionais:

$$
\begin{aligned}
&\Phi_{p,q} = \frac{\sum_{k=0}^{p} a_k \phi_k}{\sum_{k=0}^{q} b_k \phi_k}, \quad \text{a AP linear da solução exata } y;\\
&R_{p,q} = \frac{\sum_{k=0}^{p} a_k \phi_k}{\sum_{k=0}^{q} b_k \phi_k}, \quad \text{a AP não linear da solução exata } y;\\
&\Phi_{p,q}^{(N)} = \frac{\sum_{k=0}^{p} a_k^{(N)} \phi_k}{\sum_{k=0}^{q} b_k^{(N)} \phi_k}, \quad \text{a AP, ou filtro, linear de ordem } p, q \text{ de } y_N;\\
&R_{p,q}^{(N)} = \frac{\sum_{k=0}^{p} a_k^{(N)} \phi_k}{\sum_{k=0}^{q} b_k^{(N)} \phi_k}, \quad \text{a AP, ou filtro, não linear de ordem } p, q \text{ de } y_N;
\end{aligned}
$$

Relativamente aos erros das aproximações racionais:

$$
\begin{aligned}
&\delta\Phi_{p,q}^{(N)} = y - \Phi_{p,q}^{(N)}, \quad \text{o erro do filtro linear } \Phi_{p,q}^{(N)};\\
&\delta R_{p,q}^{(N)} = y - R_{p,q}^{(N)}, \quad \text{o erro do filtro não linear } R_{p,q}^{(N)};\\
&\Delta\Phi_{p,q}^{(N)} = \left|y - \Phi_{p,q}^{(N)}\right|, \quad \text{o erro absoluto do filtro linear } \Phi_{p,q}^{(N)};\\
&\Delta R_{p,q}^{(N)} = \left|y - R_{p,q}^{(N)}\right|, \quad \text{o erro absoluto do filtro não linear } R_{p,q}^{(N)}.
\end{aligned}
$$

Os erros $\delta\Phi_{p,q}^{(N)}$ e $\delta R_{p,q}^{(N)}$, com que os filtros racionais aproximam a função y, estão fundamentalmente relacionados com dois tipos de erros:

1. erros devidos ao método espectral δy_N.
2. erros devidos à aproximação de Padé $y_N - \Phi_{p,q}^{(N)}$ ou $y_N - R_{p,q}^{(N)}$.

Por sua vez estes erros estão ambos afetados por erros numéricos, devido ao uso de aritmética finita. De facto tem-se que os erros dos filtros lineares (e dos filtros não lineares[2])

$$
\delta\Phi_{p,q}^{(N)} = \underbrace{y - \Phi_{p,q}}_{\text{erro I}} + \underbrace{\Phi_{p,q} - \Phi_{p,q}^{(N)}}_{\text{erro II}}
$$

onde o erro I é devido à aproximação de Padé, e que geralmente, pelo menos com p e q suficientemente grandes, esperamos ser desprezável relativamente ao erro II. O erro II depende dos erros dos coeficientes do numerador $\delta a_k^{(N)} = a_k - a_k^{(N)}$ e do denominador $\delta b_k^{(N)} = b_k - b_k^{(N)}$ do filtro $\Phi_{p,q}^{(N)}$. Estes erros dependem dos erros

1. Erros nos coeficientes da solução espectral $\delta c_k^{(N)}$,
2. Erros numéricos no cálculo do filtro $\Phi_{p,q}^{(N)}$.

Por sua vez os erros $\delta c_k^{(N)}$ dependem dos erros devidos à projecção espectral Q_N definida por (1.20) e (1.21) e dos erros numéricos no cálculo da solução espectral y_N. As propriedades de convergência dos métodos espectrais sugerem que poderíamos menosprezar os erros $\delta c_k^{(N)}$ ao considerarmos N suficientemente grande. No entanto, tal como referido na subsecção 1.5.1, o número de condição das matrizes $\mathbf{\Gamma}_{\boldsymbol{v}}^{(\boldsymbol{N})}$ definidas em (1.54) cresce com N, perturbando fortemente as aproximações $c_k^{(N)}$. Algum progresso na estabilização destes cálculos, foi conseguido com o trabalho apresentado na subsecção 1.5.2. Esta relação existente entre os vários erros que interferem no cálculo dos filtros pode eventualmente:

- provocar a existência de pares de Froissart,
- destruir as boas propriedades dos aproximantes $\Phi_{p,q}$.

[2]Para os filtros não lineares a descrição é em tudo idêntica.

Na próxima secção iremos aplicar este método de filtragem a dois problemas lineares. No primeiro problema filtramos a solução Chebyshev-Tau de uma equação diferencial de primeira ordem com condição inicial no ponto médio do intervalo de ortogonalidade cuja solução possui um corte de bifurcação que une um ponto extremo do intervalo de ortogonalidade $z = -1$ ao ponto $z = \infty$. No segundo problema filtramos uma solução de Chebyshev-colocação, que exibe o fenómeno de Gibbs, de uma equação de segunda ordem com condições fronteira de Dirichlet.

3.3 Filtragem de problemas lineares

Exemplo 3.3.1. Neste primeiro exemplo, iremos aplicar o processo de filtragem à solução Chebyshev-Tau da equação diferencial ordinária

$$(x+1)\frac{\mathrm{d}y}{\mathrm{d}x} - \frac{1}{2}y = 0, \quad x \in]-1, 1[\tag{3.1}$$

com condição inicial $y(0) = \frac{\pi}{4}\sqrt{2}$.

Este problema tem como solução a função $y(x) = \frac{\pi}{4}\sqrt{2(x+1)}$, analítica em $\mathbb{C}\backslash]-\infty, -1]$. A função y possui um corte de bifurcação e $\zeta = -1$ é a singularidade mais próxima do intervalo $I = [-1, 1]$. Podemos observar, para este problema, que o algoritmo utilizado para calcular a solução Tau com polinómios de Chebyshev é estável até à ordem $N = 150$. Para soluções y_N com $N > 150$ as matrizes $\boldsymbol{\gamma_T}^{(N)}$ do sistema de equações linear (1.54) são mal condicionadas.

Como a solução y possui a singularidade ζ num extremo do intervalo de ortogonalidade a convergência do método Tau é lenta. Ilustramos este facto na Figura 3.1.

Podemos igualmente calcular o erro absoluto nos coeficientes espectrais, dado que é conhecida a expansão de Chebyshev da função y [Pas84],

$$y(x) = \sum_{k=0}^{\infty} c_k T_k(x) = 1 + \sum_{k=1}^{\infty} (-1)^{k+1} \frac{2}{(4k^2-1)} T_k(x).$$

Na Figura 3.2 indicamos estes erros para soluções de ordem $N = 10, 30$ e 150. Podemos observar que os valores de $\Delta c_k^{(N)}$, $k = 0, 1, \ldots, N$ diminuem quando a ordem N aumenta e que para cada valor de N os valores de $\Delta c_k^{(N)}$ decrescem com o aumento de k com a exceção do último, $k = N$, que é de uma ordem superior aos restantes. Aqui devemos salientar que este padrão é típico dos métodos espectrais aplicados a problemas diferenciais em que as funções base não satisfazem as condições suplementares. Esta observação faz com que se deva, geralmente, evitar o uso do último coeficiente espectral no processo de filtragem para os problemas com uma condição suplementar.

Tendo em vista melhorar a aproximação fornecida por y_{150} devemos escolher um bom AP. Para o efeito construimos a tabela de Froissart de ACP lineares usando os coefici-

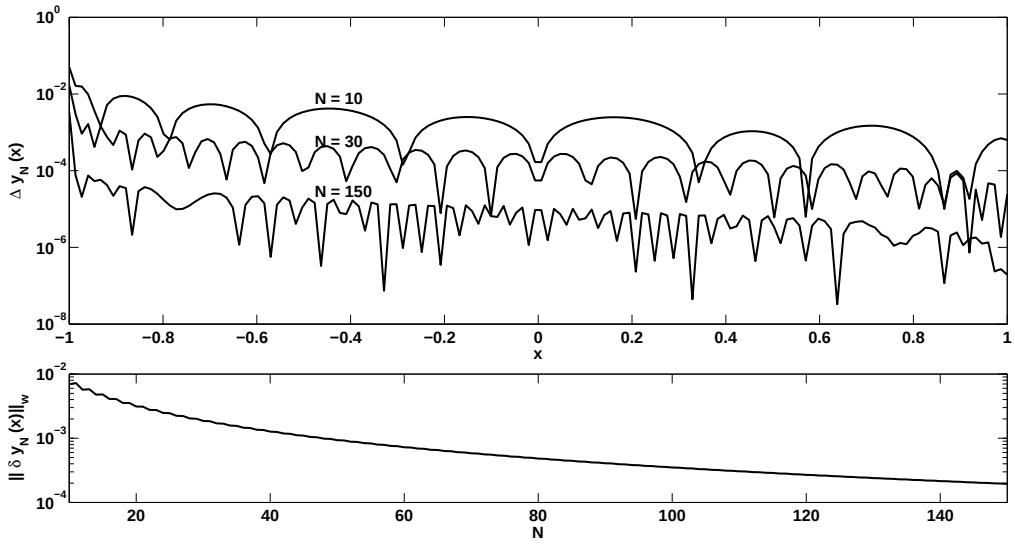


Figura 3.1: (Em cima) Erros absolutos para soluções Tau de ordem $N = 10, 30$ e 150 (Em baixo) Erro da solução Tau na norma pesada $\|\delta y_N(x)\|_w$ para valores de $N = 10, 11, \ldots, 150$.

entes $c_k^{(150)}$, e procedemos de modo análogo ao indicado na secção 2.8. Se optarmos pelo estudo da sucessão diagonal dos ACP, podemos observar na tabela de Froissart 3.3, que o aproximante $\Phi_{10,10}^{(150)}$ é o ACP de ordem mais elevada a pertencer à região livre de pares de Froissart.

Relativamente aos ACP diagonais incluídos na tabela de Froissart 3.3 e que possuem pares de Froissart, observou-se que apenas o filtro $\Phi_{14,14}^{(150)}$ não possui zeros/pólos no intervalo de ortogonalidade. Na Figura 3.4 representamos a localização de zeros/pólos de filtros diagonais.

> **Observação:** *Analogamente aos exemplos do capítulo* 2 *optamos por não incluir na Figura 3.4 todos os pólos e zeros que representam a linha de ramificação da função y para facilitar a interpretação dos dados mais relevantes. Esta nota é válida para todas as figuras, relativas à localização de pólos/zeros, existentes neste capítulo.*

Os ACP diagonais disponíveis, a partir de y_{150}, que não estão na tabela da Figura 3.3, ou seja os ACP $\Phi_{p,p}^{(150)}$ com $25 < p < 50$, possuem todos pares de Froissart no intervalo de ortogonalidade e estes pares tendem a distribuir-se por todo o intervalo de ortogonalidade quando p se aproxima de $p = 50$, ver Figura 3.5. Esta distribuição dos pares de Froissart, leva-nos a estabelecer uma relação entre a localização de pares de Froissart de ACP de séries de Chebyshev perturbadas com ruídos do tipo I, ver secção 2.7.1, e os erros introduzidos no cálculo dos filtros $\Phi_{p,p}^{(150)}$, ver subsecção 3.2, para valores de $25 < p < 50$.

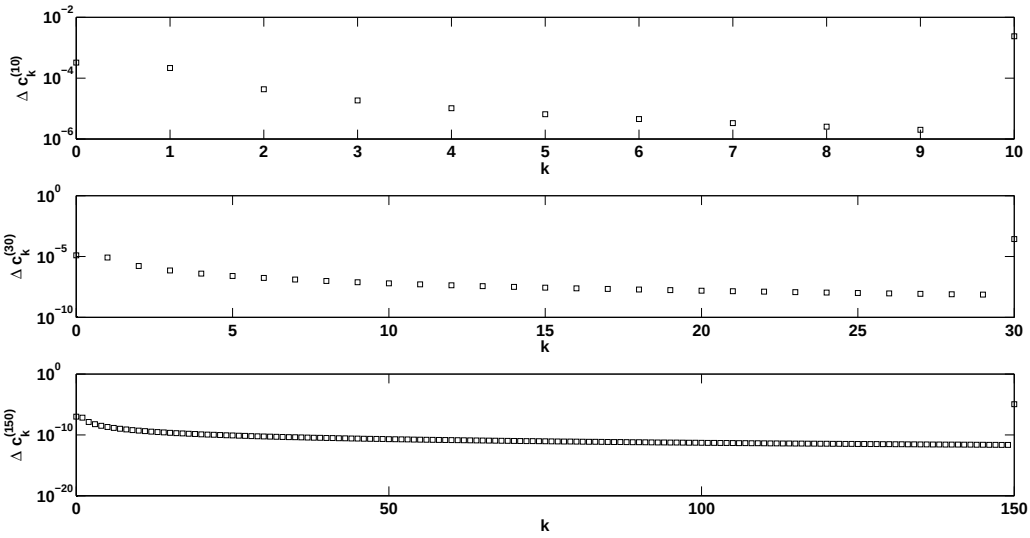


Figura 3.2: Erros absolutos $\Delta c_k^{(N)}$ dos coeficientes das soluções Tau de ordem: $N = 10$ (Em cima), $N = 30$ (No meio) e $N = 150$ (Em baixo).

Consequentemente neste exemplo (e nos exemplos seguintes) é apenas necessário estudar a localização dos pólos/zeros para AP diagonais, $\Phi_{p,p}^{(N)}$, para valores de p relativamente baixos (comparados com a ordem, N, dos aproximantes espectrais).

Para verificarmos a eficácia deste método de filtragem podemos observar, na Figura 3.6, que as aproximações obtidas após filtragem $\Phi_{10,10}^{(150)}$ e $\Phi_{14,14}^{(150)}$ melhoram efetivamente, em termos do erro absoluto $\Delta\Phi_{p,q}^{(150)}$, a aproximação dada pelo método espectral y_{150}. Além disso, o filtro $\Delta\Phi_{14,14}^{(150)}$ apenas melhora ligeiramente a aproximação dada pelo filtro $\Delta\Phi_{10,10}^{(150)}$ para valores próximos da singularidade $\zeta = -1$.

Obviamente poderiam existir filtros não diagonais que melhorassem estes resultados. Todavia, após compararmos os erros dos aproximantes $\Delta\Phi_{p,q}^{(150)}$, $p \neq q$ localizados na região livre de pares de Froissart da tabela indicada na Figura 3.3 verificamos que todos apresentam uma aproximação pior do que as aproximações diagonais $\Delta\Phi_{10,10}^{(150)}$ e $\Delta\Phi_{14,14}^{(150)}$.

Localização de singularidades

Como foi dito atrás, este método possui ainda outras vantagens, a localização de singularidades e a extensão analítica da solução espectral. Começamos por analisar os resultados obtidos na estimação da singularidade $\zeta = -1$. Note-se que a existência de uma sequência de zeros e pólos alternados, ver Figuras 3.4 e 3.5, já nos fornece a informação de que a solução do problema (3.1) possui um ramo de bifurcação. Usando as relações (2.46) e (2.47) obtivemos as estimativas para $\zeta = -1$ indicadas na Tabela 3.1, onde os valores indicados foram obtidos por arredondamento na oitava casa decimal. Os pólos dos AP

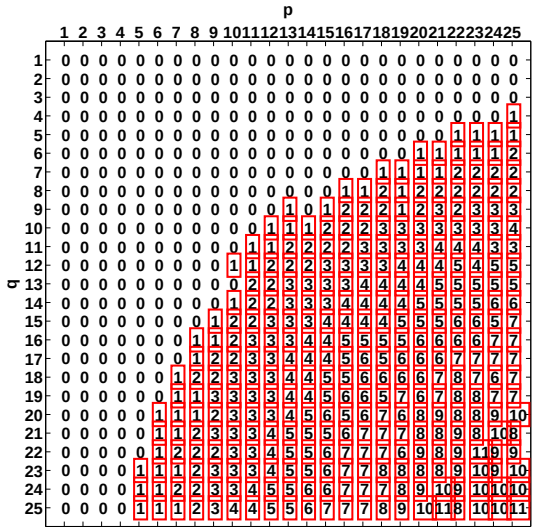


Figura 3.3: Tabela de Froissart com uma tolerância de 10^{-3} para valores de $p, q \leq 1, 2, \ldots, 25$ do exemplo 3.1.

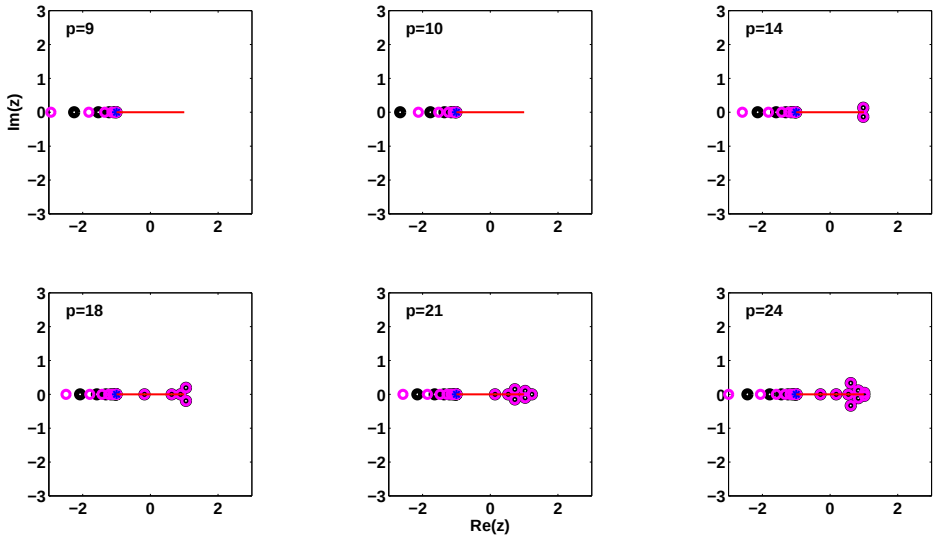


Figura 3.4: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de ACP diagonais da solução espectral y_{150} para valores de $p = 9, 10, 14, 18, 21$ e 24.

$\Phi_{p,1}^{(150)}$ dão uma primeira estimativa $x_{145,1} = -1.00014075$ mas, é possível melhorar esta estimativa usando o pólo $x_{145,2}^{+} = -1.00007608$, do aproximante $\Phi_{145,2}^{(150)}$, mais próximo da estimativa $x_{145,1} = -1.00014075$.

Relativamente à outra valência deste processo de filtragem, ou seja ao prolongamento analítico, podemos observar na Figura 3.7 o gráfico dos mesmos erros incluídos na Fi-

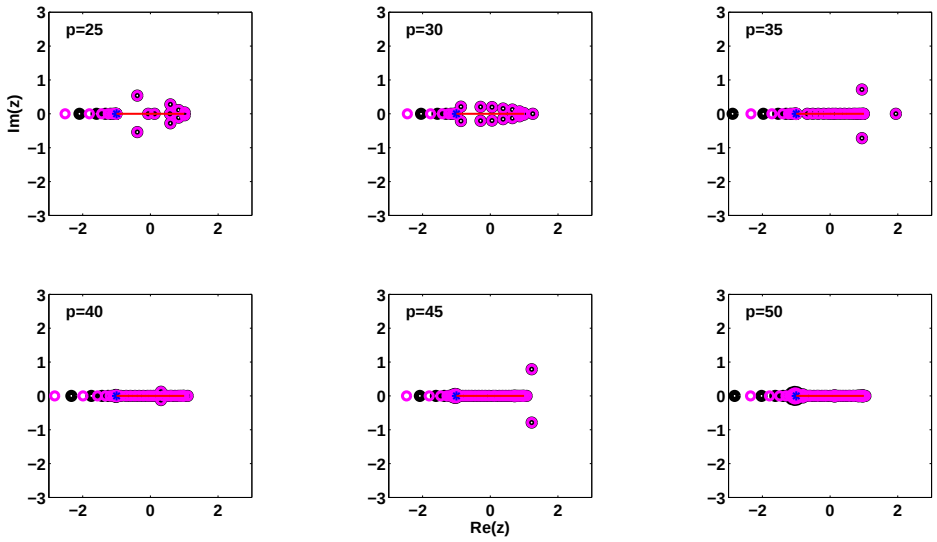


Figura 3.5: Distribuição dos zeros (círculos a magenta) e pólos (pontos a preto) de ACP diagonais, $\Phi_{p,p}^{(150)}$, da solução espectral y_{150} para valores de $p = 25, 30, 35, 40, 45$ e 50. A localização dos pares de Froissart possui um padrão similar às expansões de Chebyshev perturbadas com ruídos do tipo I.

p	5	45	85	105	145
$x_{p,1}$	-1.08888889	-1.00141928	-1.00040575	-1.00026705	-1.00014075
$x_{p,2}^{+}$	-1.042172471	-1.00075665	-1.00021841	-1.00014406	-1.00007608
$x_{p,2}^{-}$	-1.559331288	-1.00853164	-1.00245781	-1.00162066	-1.00085549

Tabela 3.1: Pólos, $x_{p,1}$, dos aproximantes $\Phi_{p,1}^{(150)}$ e pólos, $x_{p,2}^{+}$, $x_{p,2}^{-}$, de $\Phi_{p,2}^{(150)}$ para alguns valores p.

gura 3.6, no intervalo $[-1, 10]$. Podemos observar que a aproximação Tau y_{150} apenas tem significado para valores de x pertencentes ao intervalo de ortogonalidade e que as aproximações de Padé continuam a ser uma aproximação razoável para valores de x fora do intervalo de ortogonalidade embora, a aproximação perca qualidade. Pode-se ainda notar que no intervalo de ortogonalidade as funções erro $\Delta\Phi_{10,10}^{(150)}$ e $\Delta\Phi_{14,14}^{(150)}$ são quase indistinguíveis na escala do gráfico mas, para valores de x suficientemente afastados do intervalo de ortogonalidade o erro $\Delta\Phi_{14,14}^{(150)}(x)$ é claramente inferior ao erro $\Delta\Phi_{10,10}^{(150)}(x)$ em quase todos os pontos do intervalo observado.

Concluímos este exemplo com a seguinte observação relativamente à comparação dos métodos de filtragem da solução Tau via AP linear *versus* AP não linear.

Observação: *Começamos por salientar que ao comparar AP lineares com AP não lineares devemos sempre utilizar AP que usem conjuntos de dados*

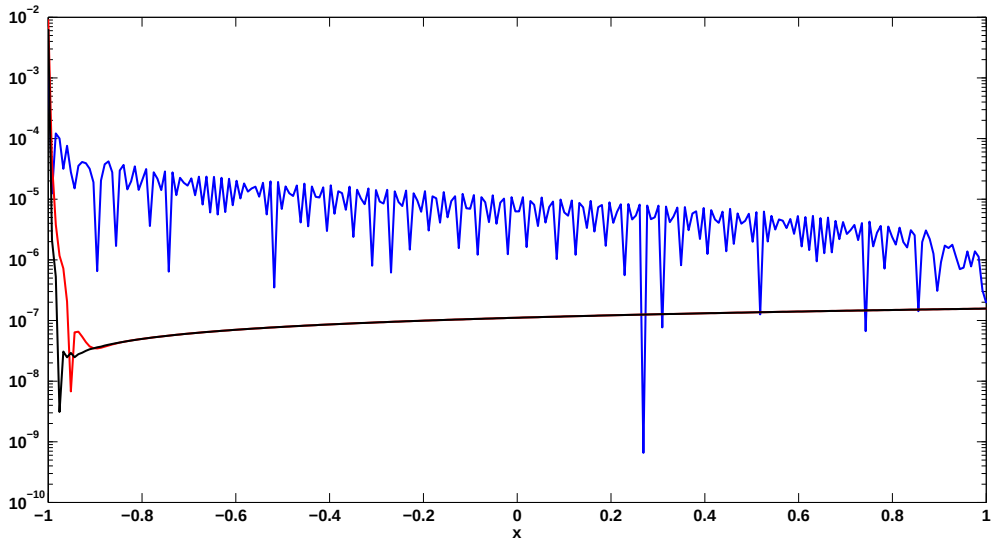


Figura 3.6: (A azul) Erro absoluto da solução Tau de ordem 150, (A vermelho) $\Delta\Phi^{(150)}_{10,10}$ (A preto) $\Delta\Phi^{(150)}_{14,14}$.

> *semelhantes. Ou seja, deve-se comparar a aproximação de um AP linear,* $\Phi_{p,q}$*, com um AP não linear* $R_{p',q'}$ *tais que* $p+2q$ *esteja próximo* $p'+q'$*. Os resultados obtidos para AP não lineares foram em tudo idênticos aos obtidos pelos AP lineares pelo que optamos não os incluir aqui. De facto, a utilização de AP não lineares não melhorou os resultados obtidos pelos AP lineares* $\Delta\Phi^{(150)}_{10,10}$ *e* $\Delta\Phi^{(150)}_{14,14}$*. Podemos explicar este comportamento com o seguinte exemplo. O cálculo dos AP* $\Phi^{(150)}_{14,14}$ *e* $R^{(150)}_{21,21}$ *apenas exige o conhecimento dos primeiros* 43 *coeficientes espectrais. Contudo a entrada* $n_{21,21}$ *da tabela de Froissart dos AP não lineares é superior à entrada* $n_{14,14}$ *da tabela de Froissart dos AP lineares e além disso possui pares de Froissart no intervalo de ortogonalidade.*

Soluções afetadas pelo fenómeno de Gibbs

O próximo exemplo serve para testar este método na filtragem a um problema cuja solução espectral exibe o chamado fenómeno de Gibbs. De entre as muitas referências existentes na literatura relacionados com este tema destacamos [Bre04] na qual o autor apresentou exemplos de filtragem de séries de Chebyshev usando o algoritmo-ϵ de Wynn, e [BMW08], na qual os autores estabeleceram uma relação entre as aproximações obtidas em [Bre04] e os ACP não lineares, e construíram uma estimativa do erro para uma classe de funções hipergeométricas. Contudo recordamos que no nosso trabalho usamos coeficientes espectrais no cálculo dos filtros, em vez dos coeficientes de Chebyshev usados nas referências atrás mencionadas.

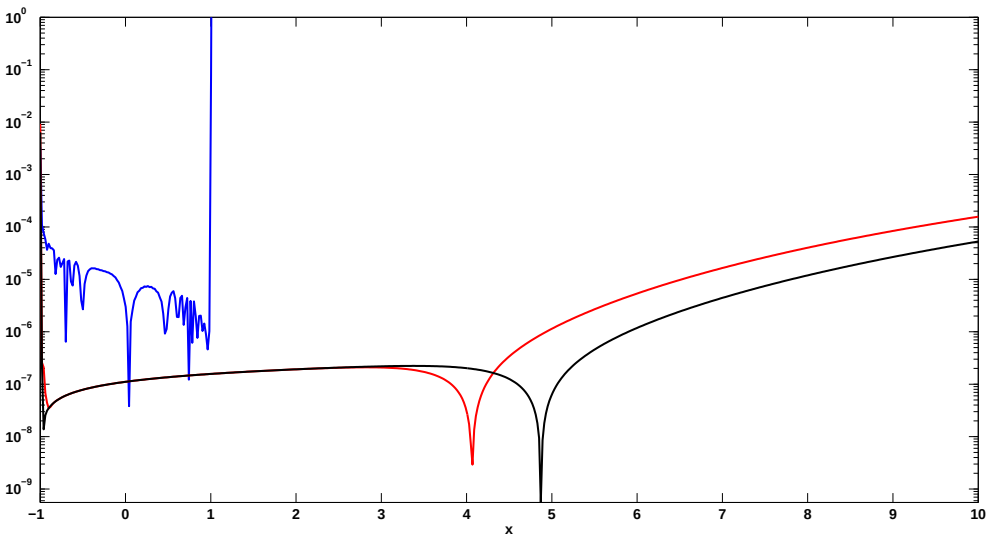


Figura 3.7: Erros absolutos, no intervalo $[-1, 10]$, da solução Tau y_{150} de ordem 150 (A azul), e dos filtros $\Phi^{(150)}_{10,10}$ (A vermelho) e $\Phi^{(150)}_{14,14}$ (A preto) do problema 3.1.

Exemplo 3.3.2. Neste exemplo consideramos a equação linear de segunda ordem

$$\epsilon \frac{\mathrm{d}^2 y}{\mathrm{d}x^2} + 2x \frac{\mathrm{d}y}{\mathrm{d}x} = 0, \quad \epsilon \in \mathbb{R}^+ \tag{3.2}$$

cuja solução geral se escreve da forma

$$y_\epsilon(x) = C_1 \sqrt{\epsilon \pi} \mathrm{erf}\left(\frac{x}{\sqrt{\epsilon}}\right) + C_2$$

onde C_1 e C_2 são constantes e erf é a *função erro*, [AS65], definida no plano complexo pelo integral

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-w^2} \mathrm{d}w.$$

A função $\mathrm{erf}(z/\sqrt{\epsilon})$ possui série de Taylor

$$\mathrm{erf}\left(\frac{z}{\sqrt{\epsilon}}\right) = \frac{2}{\sqrt{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k z^{2k+1}}{k!(2k+1)\epsilon^{2k+1}},$$

convergente para todo o $z \in \mathbb{C}$ e consequentemente $\mathrm{erf}(z/\sqrt{\epsilon})$ é uma função inteira. Na restrição à reta real e para valores de ϵ suficientemente pequenos a função $\mathrm{erf}(z/\sqrt{\epsilon})$ é um bom aproximante da função descontínua sign definida por

$$\mathrm{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}.$$

De facto, se $x \in \mathbb{R}$, tem-se $\lim_{\epsilon \to 0^+} \text{erf}(x/\sqrt{\epsilon}) = \text{sign}(x)$, dado que se $x = 0$ o integral é nulo e se x$\neq$ 0 tem-se

$$\lim_{\epsilon \to 0^+} \int_0^{x/\sqrt{\epsilon}} e^{-w^2} dw = \frac{\sqrt{\pi}}{2} \text{sign}(x).$$

Impondo à equação (3.2) as condições de fronteira de Dirichlet $y(-1) = \text{erf}(-\epsilon^{-1/2})$ e $y(1) = \text{erf}(\epsilon^{-1/2})$ então a função

$$y(x) = \text{erf}\left(\frac{x}{\sqrt{\epsilon}}\right) \tag{3.3}$$

é a solução deste problema, onde por uma questão de simplificação se deixou cair o sub índice ϵ na função y.

Iremos resolver este problema para o valor do parâmetro $\epsilon = 10^{-8}$ usando o algoritmo do método de colocação-Galerkin descrito no exemplo 1.4.1. De salientar que apesar da solução ser uma função ímpar usamos os polinómios de Chebyshev ortogonais no intervalo $I = [-1, 1]$. Esta opção justifica-se pela necessidade de se estudar o comportamento dos filtros numa vizinhança do ponto $x = 0$.

Observamos que, para este valor do parâmetro ϵ, a aproximação da função sign(x) por y apresenta erro máximo absoluto

$$\max_{x \in [-1,1]} |\text{sign}(x) - y(x)|$$

da ordem da precisão da máquina 10^{-16}. Além disso, com as nossas experiências numéricas observamos que o método de colocação apenas é estável até à ordem $N = 46$ e conforme se ilustra na imagem à esquerda da Figura 3.8, para $N = 46$ a solução de colocação é afetada pelo fenómeno de Gibbs.

Dado que não temos acesso a fórmulas fechadas dos coeficientes de Chebyshev c_k da função y, não é possível calcular os erros dos coeficientes espectrais $c_k^{(46)}$. Contudo, como são conhecidas fórmulas fechadas para os coeficientes de Chebyshev, $\tilde{c}_k$, da função sign , [Riv74],

$$\text{sign}(x) = \sum_{k=0}^{\infty} \tilde{c}_k T_k(x) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^{k+1}}{2k-1} T_{2k-1}(x).$$

estimámos os erros nos coeficientes espectrais usando o erro $\Delta c_k^{(46)} = \left|\tilde{c}_k - c_k^{(46)}\right|$. Note-se que devido à simetria da função sign(x) e à simetria dos polinómios de Chebyshev, os coeficientes $\tilde{c}_k$ são nulos para valores de k pares, e que os coeficientes espectrais pares calculados $c_{2k}^{(46)}$ estão muito próximos de zero. Consequentemente, para valores de k pares, os erros absolutos $\Delta c_k^{(46)}$ são todos de ordem inferior a 10^{-10}. Por este motivo apenas indicamos, na Figura 3.8 na imagem à direita, as estimativas dos erros absolutos dos coeficientes espectrais de ordem ímpar. Podemos observar que os erros absolutos dos coeficientes espectrais $\Delta c_k^{(46)}$ com $k = 1, 3, \ldots, 45$ tomam o seu valor máximo $\Delta c_1^{(46)} \approx 1.7e-2$ e decrescem monotonamente até atingirem o valor mínimo $\Delta c_{45}^{(45)} \approx 3.9e-4$.

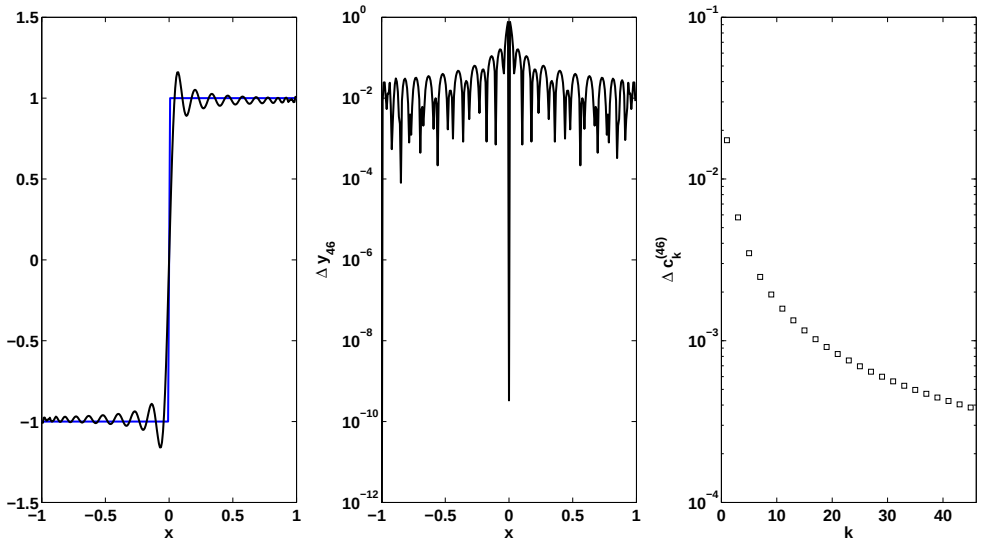


Figura 3.8: Imagem à esquerda: gráficos da função y solução do problema (3.2) (curva azul) e da solução de colocação y_{46} (curva preta). Imagem ao centro: gráfico do erro absoluto Δy_{46}. Imagem à direita: estimativa do erro $\Delta c_k^{(46)}$ dos coeficientes espectrais para valores de k ímpares.

Filtros lineares *vs* filtros não lineares

Iremos estudar o comportamento dos filtros lineares, $\Phi_{p,q}^{(46)}$, e dos filtros não lineares $R_{p,q}^{(46)}$ da função (3.3). Na Figura 3.9 apresenta-se a tabela de Froissart para os filtros lineares, com uma tolerância de 10^{-3}. As entradas $n_{p,q}$ assinaladas com um asterisco traduzem o facto de que os filtros $\Phi_{p,q}^{(46)}$ não são atingíveis com o conhecimento dos 47 coeficientes de y_{46}. Analisando esta tabela pode-se observar que na região em que os AP lineares possuem pares de Froissart todas as entradas possuem o mesmo valor $n_{p,q} = 2$. Esta propriedade revela que, se não existirem pólos no intervalo I, será possível encontrar bons filtros na região onde os AP possuem pares de Froissart. De facto o grau numérico dos numeradores, $\nu_{p,q}^N$, e dos denominadores, $\nu_{p,q}^D$, dos AP $\Phi_{p,q}$ aumentam quando aumentamos os valores de p e q porque o valor de $n_{p,q}$ é constante.

Analisando a localização dos pólos e zeros de filtros diagonais lineares, $\Phi_{p,p}^{(46)}$, e não lineares, $R_{p,p}^{(46)}$, Figura 3.11 , observamos o seguinte padrão:

1. existe um zero estável localizado perto da origem que representa o zero da solução y do problema. A distância deste zero à origem é da ordem de 10^{-10} para $p = 1$ e diminui quando p cresce.

2. existe um pólo (zero) fantasma para valores de p ímpares (pares) cuja distância à origem aumenta quando o valor de p aumenta.

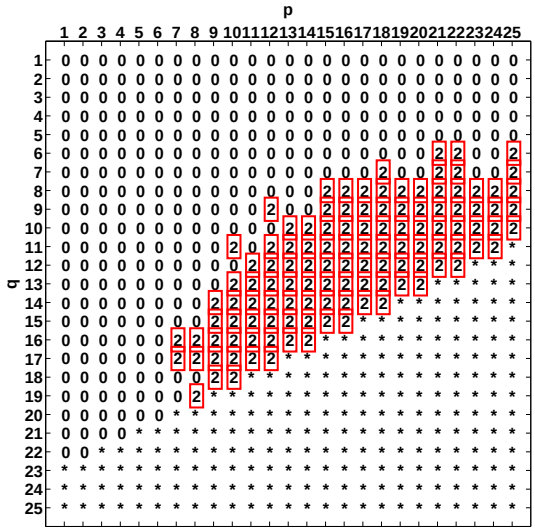


Figura 3.9: Tabela de Froissart da solução de colocação y_{46} do problema (3.2) com uma tolerância de 10^{-3}.

3. se a ordem (p, q) dos filtros, lineares ou não lineares, pertencer à região da tabela de Froissart livre de pares de Froissart todos os zeros e pólos se encontram intercalados perto do eixo imaginário. Se a ordem dos filtros pertencer à região da tabela de Froissart com dois pares de Froissart então os pares de Froissart localizam-se próximo dos extremos do intervalo de ortogonalidade I, com um par de Froissart em cada extremo. Além disso, estes pares de Froissart aproximam-se dos pontos extremos $x = -1$ e $x = 1$ quando o valor de p cresce, e podem localizar-se no interior ou no exterior de I. Os restantes pólos/zeros localizam-se intercalados perto do eixo imaginário.

Estas observações encontram-se ilustradas na Figura 3.10, onde, por uma questão de escala, não foram incluídos os zeros e pólos mais afastados da origem. A localização de um conjunto pólos/zeros situados perto do eixo imaginário mimetiza o facto da função sign ter uma descontinuidade no ponto $x = 0$ ou o facto da função $\mathrm{erf}(x/\sqrt{\epsilon})$ sofrer uma variação brusca numa vizinhança de $x = 0$.

Observação:

1. *Os resultados observados para valores de* $N < 46$ *para os filtros lineares e para os filtros não lineares são semelhantes aos resultados obtidos para o valor de* $N = 46$. *Para valores de* $N > 46$ *o esquema de colocação torna-se instável devido ao mau condicionamento das matrizes envolvidas na resolução do sistema de equações lineares* (1.41).

2. *É interessante notar que a existência da sequencia de pólos/zeros intercalados no eixo imaginário está diretamente relacionada com o facto da solução espectral exibir o fenómeno de Gibbs. Para valores de ϵ suficientemente grandes, o valor da derivada da solução de (3.2) em $x = 0$ é pequeno e y não sofre variações bruscas. Nestes casos as soluções espectrais não exibem o fenómeno de Gibbs nem existe a sequencia de pólos/zeros dos AP diagonais localizados intercaladamente no eixo imaginário. Já para valores de ϵ suficientemente pequenos, a solução espectral possui as oscilações características do fenómeno de Gibbs próximas de $x = 0$ e os seus AP diagonais possuem uma sequencia de pólos/zeros intercalados no eixo imaginário. Ilustramos esta relação na Figura 3.11. Imagens superiores: localização de pólos/zeros de uma solução "suave", $\epsilon = 1$. Imagens inferiores: localização de pólos/zeros de uma solução "não suave", $\epsilon = 10^{-3}$.*

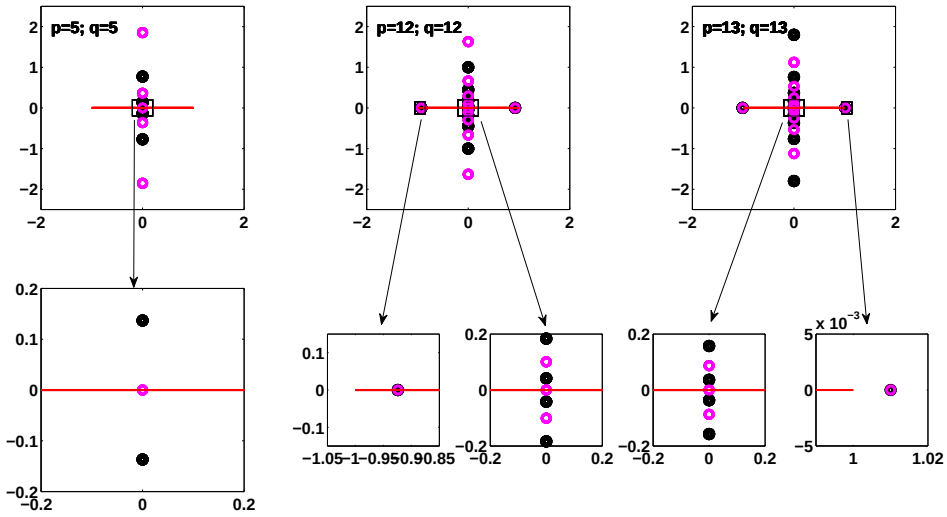


Figura 3.10: (Em cima) Localização de pólos (a preto) e zeros (a magenta) de filtros $\Phi_{p,q}^{(46)}$ da solução de colocação y_{46} do problema (3.2). (Em baixo) Ampliações das regiões delimitadas por retângulos.

Analisando a localização dos pólos e zeros de filtros diagonais observou-se:

- o filtro linear $\Phi_{13,13}^{(46)}$ não possui pares de Froissart nem pólos em I e que os filtros $\Phi_{14,14}^{(46)}$ e $\Phi_{15,15}^{(46)}$ possuem pares de Froissart em I
- o último filtro não linear disponível $R_{23,23}^{(46)}$ não possui pares de Froissart nem pólos em I.

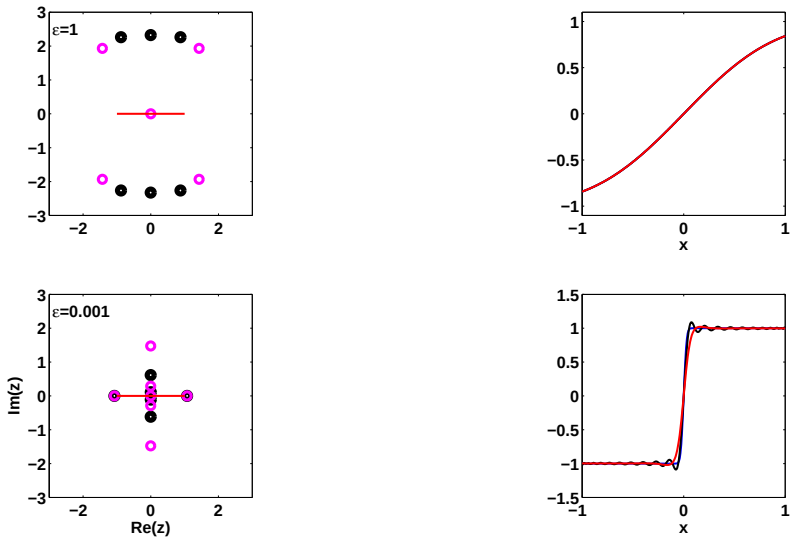


Figura 3.11: Figuras à esquerda: localização dos pólos/zeros dos filtros $\Phi_{7,7}^{(46)}$ das soluções espectrais do problema (3.2) com $\epsilon = 1$ (em cima) e com $\epsilon = 10^{-3}$ (em baixo). Figuras à direita: gráficos das funções soluções y, soluções espectral y_{46} e filtros $\Phi_{7,7}^{(46)}$ do problema (3.4) com $\epsilon = 1$ (em cima) e com $\epsilon = 10^{-3}$ (em baixo). Por uma questão de escala não incluímos os pólos e zeros mais afastados da origem.

A observação dos erros dos filtros $\Delta\Phi_{p,q}^{(46)}$ e $\Delta R_{p,q}^{(46)}$ confirmou que, de facto, os filtros $\Phi_{13,13}^{(46)}$ e $R_{23,23}^{(46)}$ são os melhores aproximantes linear e não linear respetivamente. Nas Figuras 3.12 e 3.13 ilustramos os resultados obtidos pela filtragem da solução de colocação. A Figura 3.12 contém os gráficos da solução da equação diferencial y, da solução de colocação y_{46}, do filtro linear $\Phi_{13,13}^{(46)}$ e do filtro não linear $R_{23,23}^{(46)}$. Na Figura 3.13 são ilustrados os erros absolutos da solução de colocação e dos filtros. Podemos observar que ambos os filtros reduzem o fenómeno de Gibbs, sendo que a aproximação do filtro não linear é superior à aproximação do filtro linear para valores de x próximos de zero. Salientamos que estes resultados são idênticos aos resultados obtidos no artigo [BMW08] com coeficientes exatos, no sentido que a aproximação não linear fornece uma aproximação à função sign(x) do que a aproximação linear.

Na próxima secção iremos aplicar este método de filtragem a dois problemas diferenciais não lineares. No primeiro exemplo filtramos uma solução Chebyshev-Tau de uma equação diferencial com condições fronteira e cuja solução é uma função meromorfa com um número infinito de pólos simples, sendo que um dos pólos se situa perto do intervalo de ortogonalidade. No segundo exemplo aplicamos este método de filtragem a uma solução Legendre-Tau de uma equação cuja solução tem um corte de ramificação que une um ponto de ramificação próximo de um dos extremos do intervalo de ortogonalidade ao

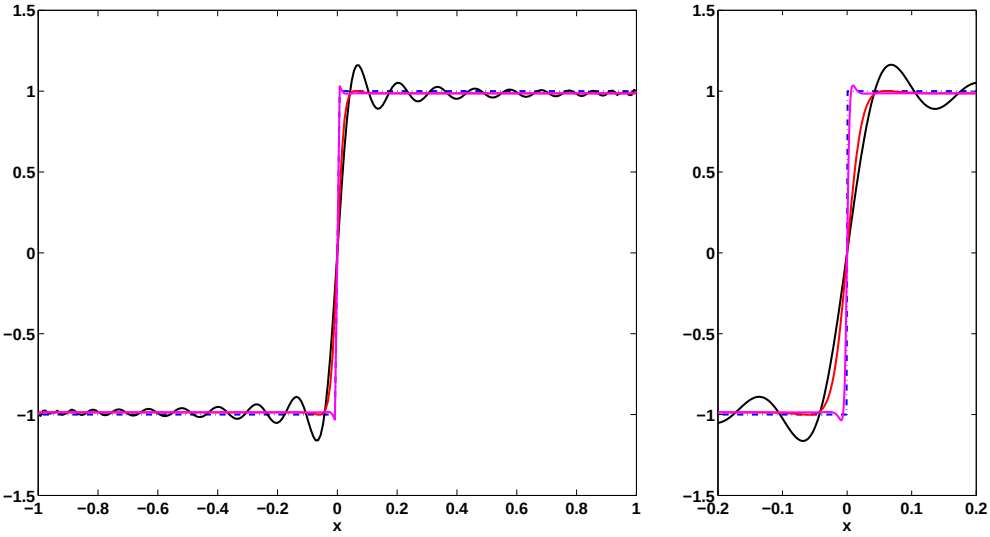


Figura 3.12: À direita, gráficos das funções solução do problema 3.2 (curva tracejada a azul), solução de colocação y_{46} (curva a preto), filtro linear $\Phi^{(46)}_{13,13}$ (curva a vermelho) e filtro não linear $R^{(46)}_{23,23}$ (curva a magenta). À esquerda, pormenor do gráfico à direita no intervalo $[-1/5, 1/5]$.

ponto de ramificação $z = \infty$.

3.4 Filtragem de soluções de problemas não lineares

No próximo exemplo estudamos o comportamento deste método de filtragem aplicado a um problema diferencial não linear com condições fronteira de Dirichlet cuja solução é uma função meromorfa com um número infinito de pólos simples.

Exemplo 3.4.1. Consideramos a equação diferencial não linear de segunda ordem

$$\epsilon \frac{\mathrm{d}^2 y}{\mathrm{d}x^2} - y \frac{\mathrm{d}y}{\mathrm{d}x} = 0, \quad x \in [0, 1], \tag{3.4}$$

onde ϵ é um número real positivo.

Esta equação sujeita às condições fronteira de Dirichlet $y(0) = 0$, $y(1) = \frac{\mathrm{tg}\alpha}{\alpha}$, com $\alpha = (2\epsilon)^{-1/2}$, tem como solução a função meromorfa,

$$y(x) = \frac{\mathrm{tg}(\alpha x)}{\alpha}$$

que possui, na reta real, um número infinito de pólos simples nos pontos

$$x_\ell = \frac{2\ell + 1}{2\alpha}\pi, \quad \ell \in \mathbb{Z} \tag{3.5}$$

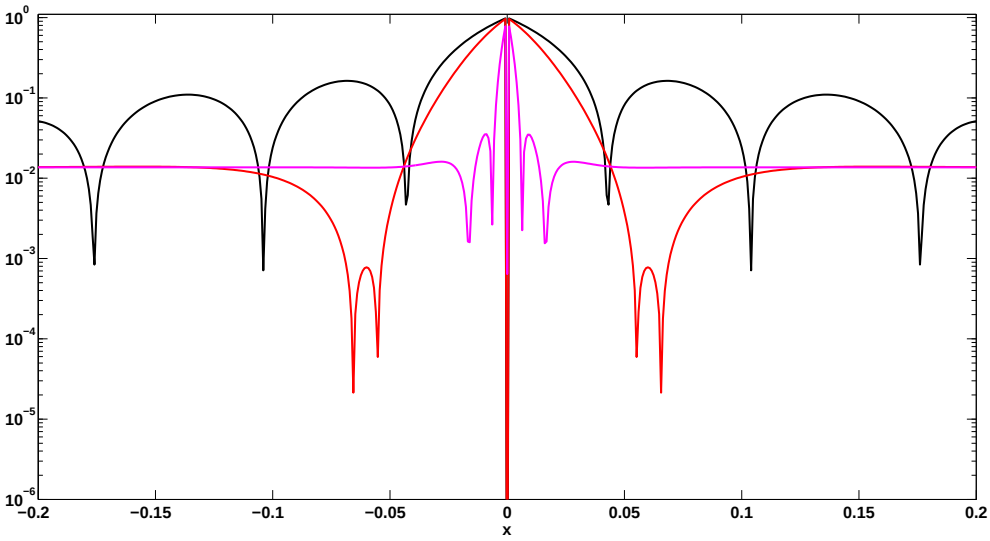


Figura 3.13: Erros absolutos da solução de colocação y_{46} (curva a preto), do filtro linear $\Phi^{(46)}_{13,13}$ (curva a vermelho) e do filtro não linear $R^{(46)}_{23,23}$ (curva a magenta) no intervalo $[-1/5, 1/5]$.

e um número infinito de zeros nos pontos

$$\tilde{x}_\ell = \frac{\ell}{\alpha}\pi, \quad \ell \in \mathbb{Z}. \tag{3.6}$$

Dado que a solução y é uma função ímpar, iremos resolver este problema aplicando o método Tau usando como funções base os polinómios de Chebyshev $T^*_k(x) \equiv T_k(2x-1)$, $k = 0, 1, \ldots$, ortogonais no intervalo $I = [0, 1]$ relativamente à função peso $w(x) = (x - x^2)^{-1/2}$ se $x \in]0, 1[$, e $w(x) = 0$ se $x \in \mathbb{R}\backslash]0, 1[$, [AS65].

Da relação (3.5), facilmente se conclui que a função y possui todos os pólos, x_ℓ, $\ell \in \mathbb{Z}$, fora do intervalo I se e somente se o parâmetro ϵ satisfizer a desigualdade $\epsilon > 2/\pi^2$. Observe-se que, para valores de $\epsilon \leq 2/\pi^2$, caso em que existem pólos de y no intervalo I, a função $y \notin L^2_w(I)$. Em face desta observação, iremos efetuar uma mudança no parâmetro $\epsilon = 2/\pi^2 + \eta$, com $\eta > 0$. Logo, para valores pequenos de η o pólo da função y mais perto do intervalo de ortogonalidade, x_0, situa-se próximo do ponto fronteira, $x = 1$, do intervalo I, e quando o valor de η cresce a singularidade x_0 afasta-se de $x = 1$. Consequentemente, a convergência do método Tau é lenta para valores pequenos de η e converge mais rapidamente à medida que o valor de η aumenta. Ilustramos este comportamento na Figura 3.14 onde apresentamos os gráficos dos erros absolutos Δy_{150} para valores de $\eta = 1/100,\ 1/500,\ 1/1000$ e $1/2000$. De referir que o método Tau, aplicado a este problema, apresenta instabilidade numérica para valores de $\eta < 1/2000$ e para ordens $N > 150$. Esta instabilidade numérica deve-se ao facto de as matrizes envolvidas apresentarem um número de condição elevado.

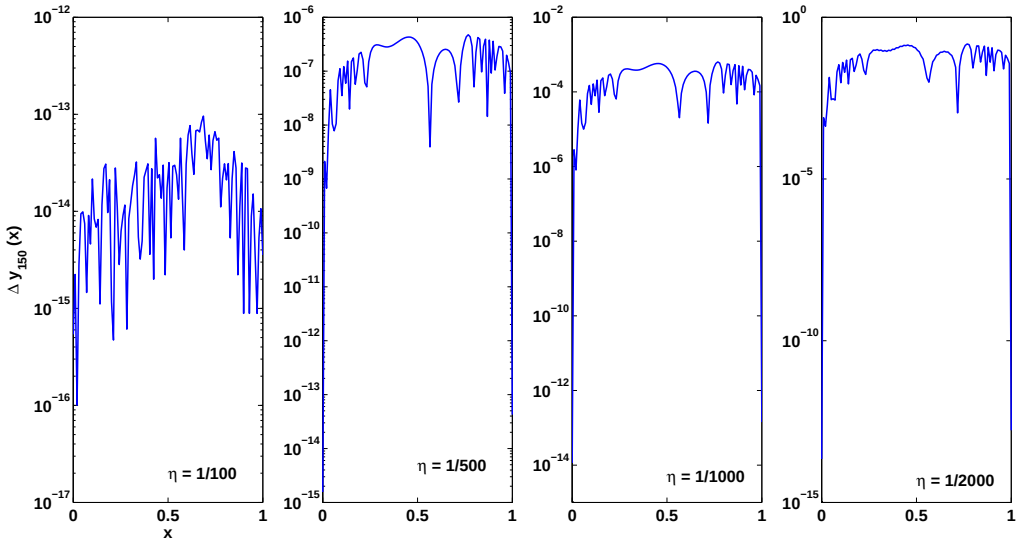


Figura 3.14: Erros absolutos da solução Tau, de ordem $N = 150$, do problema (3.4) para diferentes valores do parâmetro η.

Tendo em vista escolher um bom aproximante de Padé para filtrar as soluções Tau para os diferentes valores de η, recorremos às tabelas de Froissart. Dadas as semelhanças existentes nas tabelas de Froissart, para os diferentes valores de η, apenas apresentamos na Figura 3.15 a tabela de Froissart relativa ao valor de $\eta = 1/1000$ para valores de $p, q = 1, 2, \ldots, 25$.
Analisando os zeros e pólos dos AP diagonais, $\Phi^{(150)}_{p,p}$, observou-se que para os valores de p pertencentes à região da tabela de Froissart livre de pares de Froissart, para valores de $p \leq 6$, os pólos e zeros dos AP diagonais mimetizam a localização dos pólos e zeros da função y. Contudo, para valores de $p > 6$ todos os AP diagonais possuem pares de Froissart no intervalo de ortogonalidade. Além disso os pares de Froissart rapidamente se distribuem ao longo do intervalo de ortogonalidade, ver Figura 3.16. Este facto faz com que $\Phi^{(150)}_{6,6}$ seja o melhor aproximante diagonal disponível. O estudo das propriedades dos AP lineares revelou que todos os $\Phi^{(150)}_{p,q}$ tais que $\min(p, q) > 6$ possuem pares de Froissart no intervalo de ortogonalidade. Esta propriedade impede a existência de um AP linear que melhore a aproximação dada pelo aproximante $\Phi^{(150)}_{p,q}$. A Figura 3.17 ilustra esta propriedade e revela uma simetria entre os pólos dos AP $\Phi^{(150)}_{p,q}$ e os zeros dos AP $\Phi^{(150)}_{q,p}$.

Relativamente à filtragem da solução deste problema para valores do parâmetro $\eta = 1/100, 1/500$ e $1/2000$, obtivemos resultados idênticos aos obtidos para $\eta = 1/1000$. Por este motivo não iremos referi-los, apenas salientamos que se obteve $\Phi^{(150)}_{6,6}$, como melhor AP para todos estes os valores. Na Figura 3.18 apresentamos os erros absolutos da solução Tau e da solução Tau filtrada com filtro linear, para os valores de η referidos.

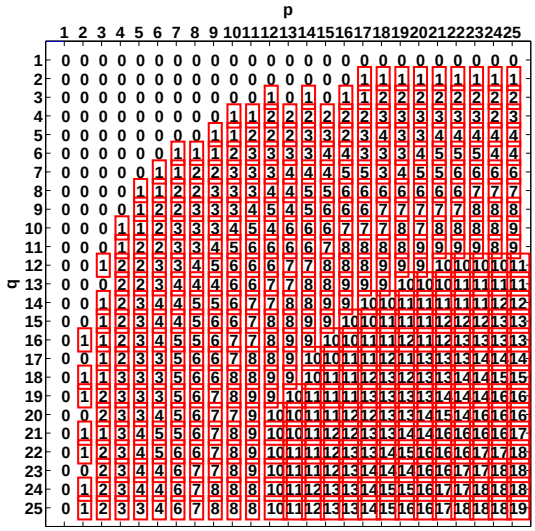


Figura 3.15: Tabela de Froissart dos AP $\Phi_{p,q}^{(150)}$ da solução Tau do problema (3.4), com $\eta = 1/1000$.

Podemos observar que o filtro apenas não melhora a aproximação espectral para o valor de $\eta = 1/100$, ou seja, quando o método Tau fornece uma aproximação com um erro próximo da precisão da máquina.

Dado que a função y possui um número infinito de pólos x_ℓ, $\ell \in \mathbb{Z}$, e que o melhor AP obtido tem como denominador um polinómio de grau 6 apenas iremos analisar a estimativa dos 6 pólos de y mais próximos do intervalo I. Definindo d_ℓ, como sendo a distância do pólo x_ℓ ao intervalo de ortogonalidade temos as seguintes desigualdades

$$d_0 < d_{-1} < d_1 < d_{-2} < d_2 < d_{-3}.$$

Indicamos na Tabela 3.2 os erros absolutos $\Delta x_\ell = |x_\ell - \hat{x}_\ell|$ para os quatro valores de η, com $\ell \in \{0, -1, 1, -2, 2, -3\}$ e onde $\hat{x}_\ell$ representa o pólo de $\Phi_{6,6}^{(150)}$ mais perto da singularidade x_ℓ. Podemos verificar que a estimativa do pólo x_0 mais próximo de I é muito precisa para os 4 valores de η e que a precisão da estimativa diminui à medida que a distância do pólo ao intervalo I aumenta. Na última linha da tabela, para $\eta = 1/2000$, as entradas com um asterisco assinalam o facto de que os pólos do AP não são números reais. Logo, considera-se que este pólos não mimetizam as 4 singularidades mais afastadas do intervalo I.

Os erros Δx_ℓ fornecem igualmente informação relevante para a análise da extensão analítica da série. De facto, os valores obtidos para os erros das estimativas das duas singularidades mais afastadas do intervalo de ortogonalidade, x_{-3} e x_2, indicam que o filtro não é um bom aproximante da solução para valores de x próximos destas singularidades.

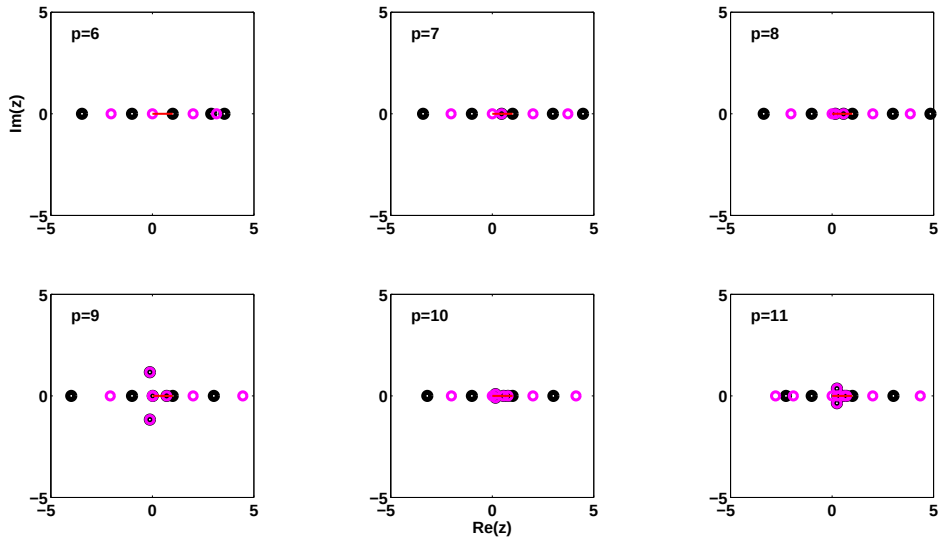


Figura 3.16: Localização dos pólos (pontos a preto) e zeros (círculos a magenta) dos AP diagonais $\Phi_{p,p}^{(150)}$ da solução Tau do problema (3.4), com $p = 6, 7, 8, 9, 10$ e 11.

	Δx_0	Δx_{-1}	Δx_1	Δx_{-2}	Δx_2	Δx_{-3}
$\eta = 1/100$	$6e-14$	$1e-5$	$6e-3$	$2e-1$	$1.4e0$	$1.1e+1$
$\eta = 1/500$	$4e-14$	$3.4e-5$	$2e-3$	$2.1e-1$	$9.6e-1$	$1.9e+1$
$\eta = 1/1000$	$1.3e-11$	$3e-5$	$2.2e-3$	$2.1e-1$	$1e0$	$1.7e+1$
$\eta = 1/2000$	$9e-10$	$7e-3$	$*$	$*$	$*$	$*$

Tabela 3.2: Erro absoluto, Δx_ℓ, das estimativas das 6 singularidades de y mais próximas do intervalo I.

Na Figura 3.19, na imagem superior, indicamos os gráficos da solução y, da aproximação Tau y_{150} e do filtro $\Phi_{6,6}^{(150)}$no no intervalo $[-4, 4]$. Na imagem inferior da mesma figura, indicamos os erros absolutos da solução Tau Δy_{150} e do filtro $\Delta\Phi_{6,6}^{(150)}$ no intervalo $[-4, 4]$, que contém as 4 singularidades de y mais próximas do intervalo I. Podemos observar que a solução Tau apenas faz sentido no intervalo de ortogonalidade $[0, 1]$ e que a qualidade da aproximação dada pelo filtro está relacionada com os erros Δx_ℓ. Por exemplo, a aproximação dada pelo filtro é claramente melhor para valores próximos da singularidade x_0 do que para valores de x próximos da singularidade x_{-1} (note-se que enquanto o erro Δx_0 é da ordem de 10^{-14} o erro Δx_{-1} é da ordem de 10^{-5}). Para as duas singularidades, no intervalo $[-4, 4]$, mais afastadas do intervalo de ortogonalidade podemos tirar conclusões análogas dado que os erros Δx_ℓ estão diretamente relacionados com as distâncias d_ℓ. Desta forma, podemos distinguir a qualidade da aproximação fornecida pelo filtro nos intervalos delimitados pelas singularidades. Como última observação,

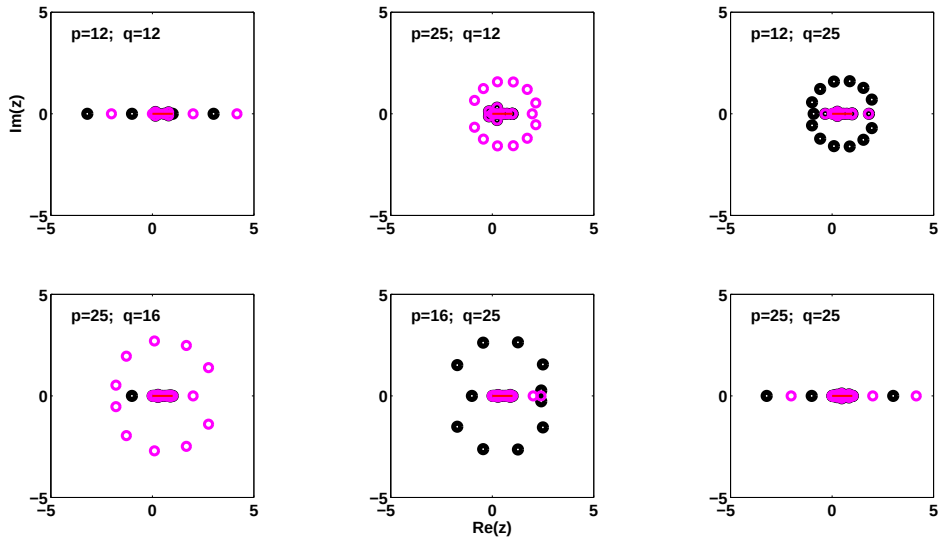


Figura 3.17: Localização dos pólos (pontos a preto) e zeros (círculos a magenta) de AP $\Phi_{p,q}^{(150)}$ e $\Phi_{q,p}^{(150)}$ da solução Tau do problema (3.4).

notamos que os AP $\Phi_{p,q}^{(150)}$, com $q > 6$ possuem todos pares de Froissart no intervalo de ortogonalidade. Este facto faz com que não seja possível melhorar as estimativas de singularidades mais afastadas do intervalo de ortogonalidade e consequentemente, melhorar a aproximação do filtro $\Phi_{6,6}^{(150)}$ em qualquer intervalo que contenha o intervalo de ortogonalidade.

Utilização de polinómios de Legendre

No próximo exemplo analisamos o comportamento deste processo de filtragem na base dos polinómios de Legendre.

Exemplo 3.4.2. Tendo em vista analisar o comportamento deste método de filtragem usando como funções base os polinómios de Legendre, consideramos a equação diferencial não linear de primeira ordem

$$\frac{\mathrm{d}y}{\mathrm{d}x} - \alpha y^3 = 0, \quad x \in]-1,1[\tag{3.7}$$

com condição $y(-1) = 1/(\alpha+1)$, onde α é um parâmetro real no intervalo $]-1,1[$. Este problema tem como solução a função

$$y(x) = \frac{1}{\sqrt{\alpha^2 + 1 - 2\alpha x}}$$

analítica em $\mathbb{C} \setminus \left[\frac{\alpha^2+1}{2\alpha}, +\infty\right]$ para $\alpha > 0$ e em $\mathbb{C} \setminus \left[-\infty, \frac{\alpha^2+1}{2\alpha}\right]$ caso $\alpha < 0$ (para $\alpha = 0$ tem-se a solução $y(x) = 1$).

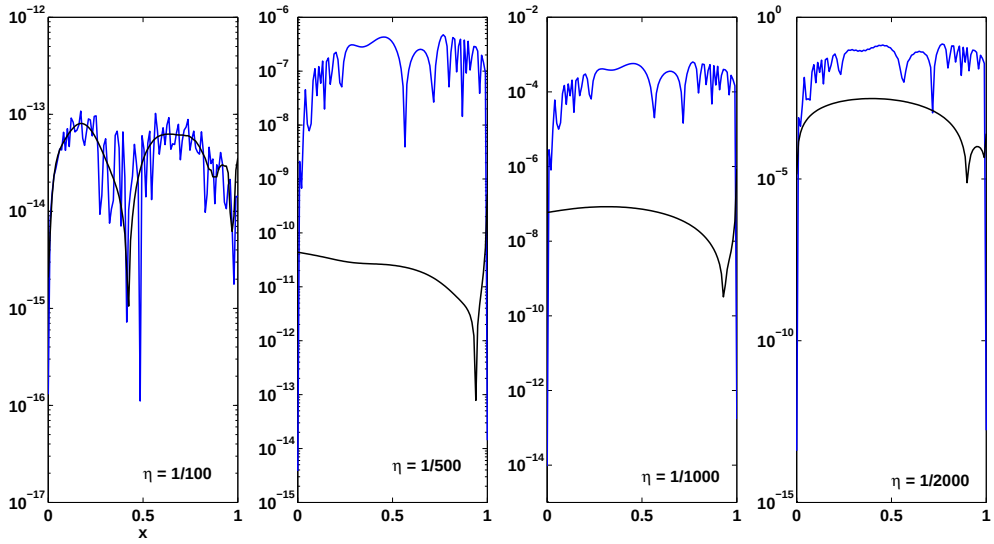


Figura 3.18: Erros absolutos Δy_{150} (a azul) e $\Delta\Phi_{6,6}^{(150)}$ (a preto) do problema (3.4).

Iremos usar o método Tau com os polinómios de Legendre, $\{P_k\}_{k\geq 0}$ ortogonais no intervalo $I = [-1, 1]$ relativamente à função peso $w(x) = 1$, como funções base para resolver o problema (3.7). Notamos que para todo α não nulo o ponto $\zeta = \frac{\alpha^2+1}{2\alpha}$ é a singularidade da função y mais próxima do intervalo de ortogonalidade, e localiza-se à direita (esquerda) do intervalo I se $\alpha < 0$ ($\alpha > 0$). Quando α tende à esquerda para 1 (à direita para -1) tem-se que a singularidade ζ tende à direita para 1 (à esquerda para -1), respetivamente. Mais, quando α tende à esquerda para 0 (à direita para 0) tem-se que a singularidade ζ tende para $+\infty$ ($-\infty$), respetivamente. Consequentemente, o método Tau terá convergência lenta para valores de $|\alpha|$ próximos de um e terá convergência rápida para valores de $|\alpha|$ próximos de zero.

É conhecida a série de Legendre da função y [AS65],

$$y(x) = \sum_{k=0}^{\infty} c_k P_k(x) = 1 + \sum_{k=1}^{\infty} \alpha^k P_k(x).$$

Fixando o valor do parâmetro $\alpha = 9/10$, tem-se que o valor da singularidade mais próxima de I é $\zeta = 1.005(5)$. Na Figura 3.20 apresentamos o gráfico do erro absoluto da solução Tau de ordem $N = 60$, Δy_{60} (imagem de cima) e o gráfico dos erros absolutos dos coeficientes $\Delta c_k^{(60)} = |\alpha^k - c_k^{(60)}|$, $k = 0, \ldots, 60$ (imagem de baixo).

Podemos observar que os erros dos coeficientes $\Delta c_k^{(60)} = |\alpha^k - c_k^{(60)}|$ são da ordem de 10^{-2} para valores de k pequenos, crescem até atingir valores da ordem de 10^{-1} para valores de k próximos de $k = 10$ e decrescem monotonamente até à ordem 10^{-3} até atingir o seu valor mínimo para $k = 60$.

Uma análise do erro Δy_{60} revela que este erro possui dois comportamentos distintos

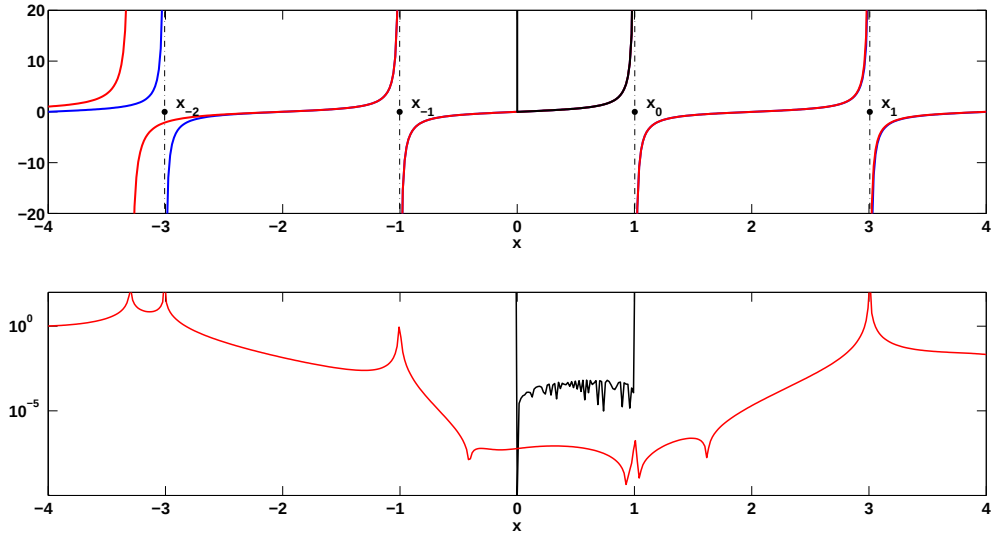


Figura 3.19: Extensão analítica do filtro $\Phi_{6,6}^{(150)}$ da solução Tau do problema (3.4). (Em cima) Gráficos das funções solução do problema y, com o valor de $\eta = 1/1000$, (linha azul), solução Tau y_{150} (linha preta) e da solução filtrada $\Phi_{6,6}^{(150)}$ (linha vermelha). (Em baixo) Gráficos dos erros absolutos Δy_{150} (linha preta) e $\Delta\Phi_{6,6}^{(150)}$ (linha vermelha).

no intervalo $[-1, 1]$. Mais exatamente, existe um ponto $\xi \in]-1, 1[$ para o qual $\Delta y_{60}(x)$ atinge valores máximos da ordem de 10^{-8} para $x < \xi$ e para valores de $x > \xi$, os valores máximos de $\Delta y_{60}(x)$ crescem exponencialmente e atingem o seu valor máximo, da ordem de 10^0, no extremo do intervalo de ortogonalidade, $x = 1$, mais próximo da singularidade ζ. Notamos que o valor de ξ não é fácil de calcular, para efeitos práticos podemos estimar o seu valor como sendo $\xi \approx 0.68$.

Tendo em vista encontrar um bom filtro para a solução Tau y_{60}, apresentamos na Figura 3.21 a tabela de Froissart com entradas $n_{p,q}$, onde $1 \leq p, q \leq 25$ e $p + 2q \leq 60$. De salientar que, ao contrário dos exemplos anteriores, a região com entradas $n_{p,q}$ positivas não é conexa. Analisando os ALP $\Phi_{p,q}^{(60)}$ livres de pares Froissart, escolhemos $\Phi_{7,7}^{(60)}$ e $\Phi_{16,22}^{(60)}$ como "candidatos" ao melhor filtro. A escolha de $\Phi_{7,7}^{(60)}$ é de certa forma óbvia e deve-se à utilização da mesma estratégia usada nos exemplos anteriores. De facto, $\Phi_{7,7}^{(60)}$ é o último AP diagonal que está na região livre de pares de Froissart e não possui pólos no intervalo de ortogonalidade, ver Figura 3.22. A escolha do outro "candidato" deve-se ao facto de $\Phi_{16,22}^{(60)}$ também pertencer à região livre de pares de Froissart, não possuir pólos no intervalo de ortogonalidade, ver Figura 3.22, e de entre os AP, $\Phi_{p,q}^{(60)}$, que satisfazem estas duas propriedades, o valor de q é máximo.

Também optámos por estudar os aproximantes $\Phi_{p,q}^{(60)}$ que verificassem as seguintes condições: o valor de $p + 2q$ ser máximo e não possuírem pólos nem pares de Froissart

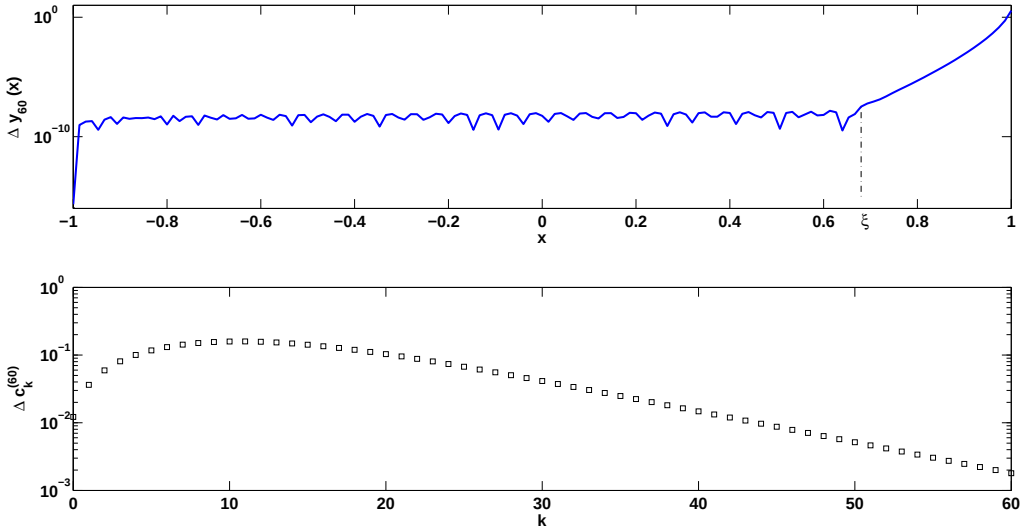


Figura 3.20: Erro absoluto da solução Tau de ordem $N = 60$, $\Delta y_{60} = |y - y_{60}|$, (Cima) e Erro absoluto nos coeficientes, $\Delta c^{(60)_k} = |\alpha^k - c_k^{(60)}|$ (Baixo) do problema (3.7) com $\alpha = 9/10$.

no intervalo de ortogonalidade. Note-se que o cálculo destes candidatos, utiliza o maior número de coeficientes espectrais. Nos testes efetuados, detetou-se que o valor máximo é $p + 2q = 57$, sendo que o AP diagonal, $\Phi_{19,19}^{(60)}$, é o único que verifica as condições acima referidas, ver Figura 3.22.

Relativamente aos erros absolutos, $\Delta\Phi_{7,7}^{(60)}$, $\Delta\Phi_{16,22}^{(60)}$ e $\Delta\Phi_{19,19}^{(60)}$ observamos as seguintes características:

1. Todos os erros, $\Delta\Phi_{7,7}^{(60)}(x)$, $\Delta\Phi_{16,22}^{(60)}(x)$ e $\Delta\Phi_{19,19}^{(60)}(x)$ têm máximos, no intervalo I, com a mesma ordem de grandeza, sendo que o gráfico de $\Delta\Phi_{16,22}^{(60)}(x)$ é praticamente coincidente com o gráfico de $\Delta\Phi_{19,19}^{(60)}(x)$,

2. No intervalo $[-1, \xi]$ onde o erro da solução Tau, Δy_{60}, é estável (com valores máximos da ordem de 10^{-8}) os erros dos três filtros são da ordem de 10^{-12} para valores de x próximos do extremo inferior do intervalo I e crescem suavemente até atingirem valores de ordem igual à do erro Δy_{60} no ponto $x = \xi$.

3. No intervalo onde o erro da solução Tau, Δy_{60}, cresce exponencialmente, $\xi < x < 1$, todos os erros são praticamente coincidentes.

Estas observações encontram-se ilustradas na Figura 3.23.

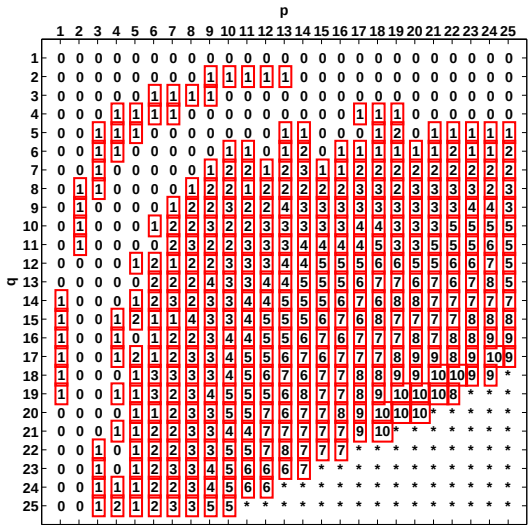


Figura 3.21: Tabela de Froissart da solução y_{60}, do problema (3.7) com $\alpha = 9/10$, com uma tolerância de 10^{-3}. As entradas não disponíveis estão assinaladas com um asterisco.

Estimativa da singularidade mais próxima do intervalo I

Analisamos de seguida a estimativa da singularidade mais próxima do intervalo I. Pode-se usar a relação (2.51) para estimar os pólos da sequência de ALP $\{\Phi_{p,1}\}_{p\geq 1}$. Lembramos que o gráfico dos filtros nos pontos perto da singularidade ζ é semelhante ao gráfico da solução Tau. Este comportamento indica que a estimativa de ζ não deverá ter a mesma qualidade das estimativas das singularidades, obtidas nos exemplos anteriores. Podemos confirmar esta observação analisando os valores dos pólos dos aproximantes $\Phi_{p,1}^{(60)}$ indicados na Tabela 3.3, arredondados na quarta casa decimal. De facto, quando o valor de p cresce, o erro absoluto das estimativas, $\Delta x_{p,1} = |\zeta - x_{p,1}|$ diminui até atingir o seu valor mínimo em $p = 9$.

p	1	8	9	10	20
$x_{p,1}$	1.1372	1.0311	1.0309	1.0312	1.0412
$\Delta x_{p,1}$	$1.316e-1$	$2.556e-2$	$2.538e-2$	$2.561e-2$	$3.564e-2$

Tabela 3.3: Estimação da singularidade mais próxima do intervalo de ortogonalidade da solução do problema (3.7), $\zeta = 1.005(5)$ usando os pólos $x_{p,1}$ dos aproximantes $\Phi_{p,1}^{(60)}$ para valores $p = 1, 8, 9, 10$ e 20.

Relativamente ao comportamento dos filtros para valores de $x < -1$, podemos observar, ver Figura 3.22, que os filtros $\Delta\Phi_{16,22}^{(60)}$ e $\Delta\Phi_{19,19}^{(60)}$ possuem pólos no intervalo $x < -1$. Deste modo apenas iremos analisar o comportamento do filtro $\Delta\Phi_{7,7}^{(60)}$. Como se pode

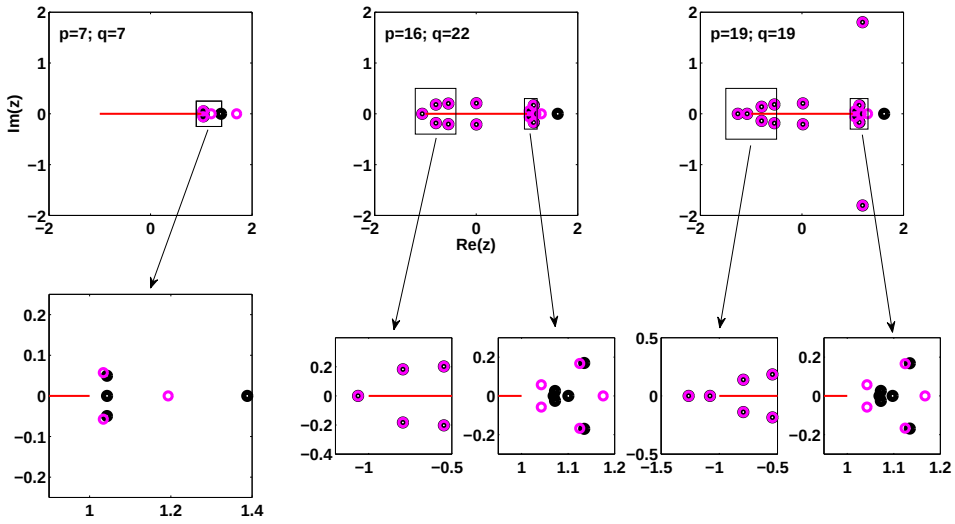


Figura 3.22: (Cima) Localização dos pólos (pontos a preto) e zeros (círculos a magenta) dos AP: $\Phi_{7,7}^{(60)}$, $\Phi_{16,22}^{(60)}$ e $\Phi_{19,19}^{(60)}$. (Baixo) Ampliação das regiões das Figuras da linha superior delimitadas por retângulos a preto.

observar na Figura 3.24 a aproximação Tau, y_{60} apenas fornece uma boa aproximação da solução do problema (3.4) no intervalo de ortogonalidade dos polinómios da base. A aproximação dada pelo filtro $\Phi_{7,7}^{(60)}(x)$ além de melhorar a aproximação dada pela solução Tau no intervalo $[-1, \xi]$, permite ainda a extrapolação no domínio da função y. Note-se que para valores de $x < -1$ o erro cresce lentamente, quando x decresce, atingindo no intervalo $[-100, 1]$ o valor máximo da ordem de 10^{-2}, no ponto $x = -100$.

3.5 Observações e conclusões

Nos exemplos apresentados neste capítulo podemos concluir que os filtros mantêm as "boas propriedades" dos aproximantes de Padé de séries generalizadas de Fourier. Ou seja, melhoram as aproximações dadas pelas soluções espectrais, permitem a localização de singularidades e fornecem uma extensão analítica das soluções espectrais.

A principal contrapartida deste método de filtragem reside no facto dos filtros de Padé $\Phi_{p,q}^{(N)}$ e $R_{p,q}^{(N)}$ possuírem pares de Froissart distribuídos ao longo do intervalo de ortogonalidade. Contudo esta contrapartida apenas se manifesta para valores de p e q relativamente elevados. De facto, nas nossas experiências, verificámos que a parcela dos erros, nos coeficientes espectrais, devido à projeção espectral não é a responsável pela existência de pares de Froissart localizados no intervalo de ortogonalidade. Note-se que os erros de projecção não são aleatórios, dependem essencialmente do operador

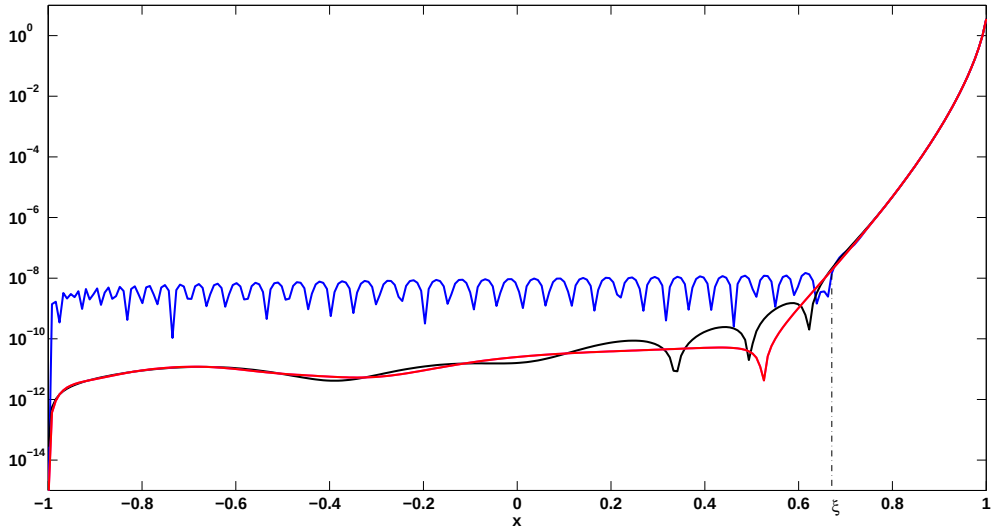


Figura 3.23: Erros absolutos, da solução Tau y_{60} (curva azul) e do filtro $\Phi^{(60)}_{7,7}$ (curva preta) e dos filtros $\Phi^{(60)}_{16,22}$ e $\Phi^{(60)}_{19,19}$ que são representados ambos pela mesma curva (vermelha), do problema (3.7) com $\alpha = 9/10$.

diferencial $\mathscr{L}$ e das funcionais lineares $\mathcal{B}$ que definem as condições fronteira de um dado problema diferencial (1.1)-(1.2) que pretendemos estudar. Portanto a existência de pares de Froissart, localizados no intervalo de ortogonalidade, deve-se ao uso de software com aritmética de precisão finita. Ou seja, os erros numéricos cometidos no cálculo da solução espectral e no cálculo dos filtros de Padé são os responsáveis pela ocorrência de pares de Froissart localizados no intervalo de ortogonalidade. Esta observação é suportada pelo facto de que os pares de Froissart surgem em filtros calculados resolvendo sistemas de equações lineares mal condicionados. Ou seja as entradas $n_{p,q}$ não nulas ocorrem, na tabela de Froissart, quando o número de condição das matrizes envolvidas no cálculo de um filtro é muito elevado.

Para os aproximantes de Padé de séries de potências existem algoritmos, via decomposição em valores singulares, que permitem remover os pares de Froissart, fornecendo os chamados *aproximantes de Padé robustos* [GGT13].

Para calcular filtros de Padé que partilhem as boas características dos AP robustos de séries de potências apresentamos aqui um conjunto de regras empíricas que permitem, geralmente, encontrar um "bom" filtro e ultrapassar, parcialmente, as contrariedades acima referidas. Estas regras foram estabelecidas com base nos resultados de todas as experiências efetuadas e aplicam-se a filtros lineares e a filtros não lineares.

Regras de procedimento na filtragem:

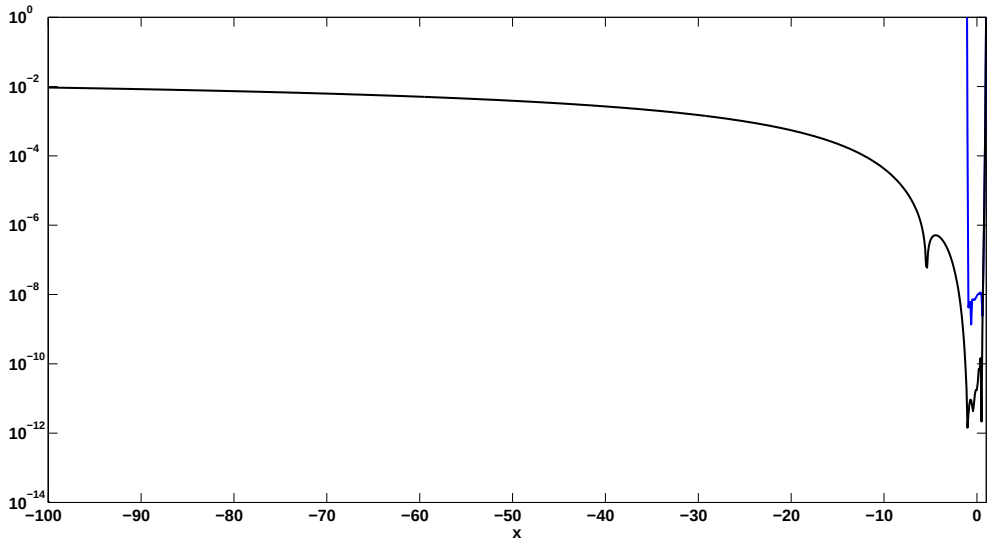


Figura 3.24: Erros absolutos, da solução Tau y_{60} do problema (3.7) (curva azul) e do filtro $\Phi_{7,7}^{(60)}$ (curva preta) no intervalo $[-100, 1]$, com $\alpha = 9/10$.

1. Encontrar uma solução espectral $y_{N_0} = \sum_{k=0}^{N_0} c_k^{(N_0)} \phi_k$, tal que os coeficientes das soluções espectrais de ordem inferior a N_0 satisfaçam a seguinte condição

$$|c_k^{(\ell)} - c_k^{(\ell-1)}| < |c_k^{(\ell-1)} - c_k^{(\ell-2)}|, \tag{3.8}$$

para todo $\ell \leq N_0$ e $0 \leq k \leq \ell - 2$. Esta condição evita, geralmente, a introdução de ruídos desnecessários nos coeficientes espectrais a utilizar no processo de filtragem.

Se escolhermos o valor de N_0 de forma a que condição (3.8) se verifique e não se verifique para $N_0 + 1$, então este critério coincide geralmente com o melhor aproximante espectral que se pode calcular. Ou seja a partir da ordem N_0 o algoritmo usado para calcular y_N é numericamente instável para valores de $N > N_0$. Esta condição é especialmente relevante quando filtramos soluções de problemas não lineares. De facto este critério é semelhante ao proposto por Ortiz [OS81], onde se resolve iterativamente um conjunto de equações lineares utilizando-se como condição de paragem do processo iterativo

$$\max_{x \in I} |y_N(x) - y_{N-1}(x)| < \text{tol}. \tag{3.9}$$

Consequentemente, tendo em vista a filtragem da solução espectral, a condição (3.8) é preferível como método de paragem à condição (3.9).

2. O próximo passo consiste em construir a tabela de Froissart dos filtros disponíveis, isto é os filtros de ordem (p, q) tais que $p + 2q \leq N_0$. Deve-se ter em atenção que a tabela deve ser estável, ou seja, as entradas $n_{p,q}$ não devem variar muito quanto alteramos o valor da tolerância, geralmente é suficiente testar para valores da tolerância entre 10^{-4} e 10^{-3}. Se determinarmos a tabela de Froissart com valores da tolerância de ordem superiores a 10^{-3} existe a possibilidade de estarmos a identificar pares de pólos e zeros que estão próximos mas que contribuem para a qualidade da aproximação do filtro.

3. Calcular uma sequência de filtros $\Phi_{p,q}^{(N_0)}$, na região da tabela de Froissart com entradas $n_{p,q}$ nulas. Geralmente é suficiente escolher uma sucessão de filtros onde os valores de p e q cresçam, p.ex. uma sucessão diagonal ou para-diagonal.

4. Identificar os seguintes padrões na localização dos pólos e zeros da sequência de filtros $\Phi_{p,q}^{(N_0)}$:

   - Pólos estáveis que correspondem geralmente a singularidades da solução do problema.
   - Um conjunto de pólos e zeros intercalados. Então a singularidade é um ponto de ramificação e a solução possui um corte de ramificação representada pela localização deste conjunto de pólos/zeros.
   - Zeros estáveis que correspondem geralmente a zeros da solução do problema.
   - Conjuntos de pólos/zeros localizados intercaladamente numa reta vertical, ou seja com parte imaginária constante, corresponde a um ponto de descontinuidade.

5. Considerar como candidato a melhor filtro o aproximante $\Phi_{p,p}^{(N_0)}$ diagonal tal que $n_{p,p} = 0$ e o valor de p é máximo.

6. Pode-se, eventualmente, encontrar melhores filtros na região da tabela de Froissart com entradas $n_{p,q}$ positivas. Tais candidatos serão os filtros $\Phi_{p,q}^{(N_0)}$ para os quais a diferença $\min\{p, q\} - n_{p,q}$ é máxima e que não possuam pólos no interior do intervalo de ortogonalidade. Note-se que para filtros diagonais a diferença $\min\{p, q\} - n_{p,q}$ coincide com o grau numérico $\nu_{p,q}$. Nas nossas experiências verificamos que a melhoria da aproximação obtida por tais filtros não é significativa relativamente aos filtros calculados no ponto anterior. Além disso, os pares de Froissart tendem a localizar-se no intervalo de ortogonalidade para valores de p e q relativamente elevados o que torna inútil a procura de filtros, nesta região, de ordens (p, q) elevadas.

Tendo em vista estimar a singularidade, da solução de um dado problema, mais próxima do intervalo de ortogonalidade, pode-se encontrar uma primeira estimativa usando os pólos dos filtros $\Phi_{p,1}^{(N)}$. Porém, devemos ter o cuidado de não usar os últimos coeficientes espectrais da solução espectral y_N. O número $N - k_0$ de coeficientes espectrais $c_k^{(N)}$, $k = 0, 1, \ldots, N - k_0$ que devemos usar, para estimar a singularidade, depende do número de condições que o problema diferencial possui. Por exemplo, no caso de o problema diferencial ter uma condição inicial, problema de Cauchy, tem-se $k_0 = 1$, e, caso o problema diferencial tenha condições fronteira de Dirichlet tem-se $k_0 = 2$. O seu cálculo pode efetuar-se usando a relação (2.45). No caso dos polinómios de Chebyshev e dos polinómios de Legendre, a relação (2.45) toma a forma (2.46), (2.49), respetivamente. Se pretendermos melhorar a primeira estimativa dada por (2.45) pode-se calcular os pólos de filtros $\Phi_{p,q}^{(N)}$, $q > 1$, e usar os pólos mais próximos da primeira estimativa como novos estimadores da singularidade. Geralmente, nas nossas experiências, verificámos que estes estimadores melhoram o resultado dado por (2.45). No caso dos polinómios de Chebyshev ortogonais no intervalo $[-1, 1]$, pode-se usar a relação (2.47) para obter os pólos dos filtros $\Phi_{p,2}^{(N)}$. Para os polinómios de Chebyshev ortogonais no intervalo $[a, b]$ pode-se usar as relações (2.46) e (2.47) para calcularmos os pólos dos filtros $\Phi_{p,1}^{(N)}$ e $\Phi_{p,2}^{(N)}$, respetivamente, tendo o cuidado de efetuar a transformação linear $\xi = 1/2((b-a)\eta + a + b)$. Este método pode igualmente ser usado para estimar eventuais singularidades, mais afastadas do intervalo de ortogonalidade, como efetuamos no exemplo 3.4.1. Ou seja, determinamos numericamente os pólos dos filtros $\Phi_{p,q}^{(N)}$, $q > 2$.

Capítulo 4

Aproximação de Padé multidimensional

O conceito de AP de séries de potências a duas variáveis foi introduzido por J. Chisholm em [Chi73] e generaliza o conceito de AP de funções dadas por uma série de potências a uma variável. Esta generalização foi posteriormente [CM74] expandida a séries de potências com um número arbitrário variáveis. Em [Mat07] A.C. Matos introduziu vários AP de séries ortogonais multidimensionais.

Neste capítulo começamos por descrever algumas dificuldades que surgem quando passamos da AP unidimensional à AP multidimensional. Posteriormente resumimos os AP de séries ortogonais multidimensionais sugeridos em [Mat07] e descrevemos os algoritmos usados neste trabalho, para calcular aproximantes de Padé bidimensionais de funções representadas por séries de Chebyshev (ACP). Finalmente iremos aplicar os ACP bidimensionais para filtrar soluções diferenciais parciais em duas variáveis e analisar os resultados obtidos.

4.1 Da AP unidimensional à AP multidimensional

A AP de séries de potências com múltiplas variáveis possui essencialmente duas dificuldades. A primeira dificuldade é de carácter conceptual [Cuy99] e a segunda relaciona-se com a convergência de sucessões de aproximantes racionais. Para descrevermos as dificuldades conceptuais começamos por recordar a definição de aproximante de Padé de séries de potências a uma variável e descrever as várias generalizações multidimensionais.

Seja f uma função representada por uma série de potências $f(x) \sim \sum_{i=0}^{\infty} c_i x^i$ e o seu aproximante de Padé $[n/m]_f(x)$. Dados dois polinómios $p(x) = \sum_{i=0}^{n} a_i x^i$ e $q(x) =$

$\sum_{i=0}^{m} b_i x^i$ que satisfazem a condição

$$(fq - p)(x) = \sum_{i=n+m+1}^{\infty} d_i x^i, \tag{4.1}$$

tem-se que $[n/m]_f(x)$ é a forma irredutível de $p(x)/q(x)$. A condição (4.1) conduz ao sistema de $n + m + 1$ equações lineares

$$\sum_{j=0}^{i} c_j b_{i-j} - a_i = 0, \quad i = 0, \ldots, n \tag{4.2a}$$

$$\sum_{j=0}^{i} c_j b_{i-j} = 0, \qquad i = n+1, \ldots, n+m \tag{4.2b}$$

nas $n + m + 2$ incógnitas $a_i, \quad i = 0, \ldots, n$, $b_i, \quad i = 0, \ldots, m$. Dado que as equações (4.2b) formam um sistema homogéneo com m equações a $m + 1$ incógnitas é sempre possível atribuir um valor não nulo a uma das incógnitas $b_i, \quad i = 0, \ldots, m$ e tem-se sempre uma solução não trivial para o sistema (4.2b). Esta atribuição não afeta a função racional p/q uma vez que os polinómios são determinados a menos de um fator multiplicativo arbitrário. Para garantir a existência do aproximante de Padé $[n/m]_f$ para todos os inteiros não negativos n e m, teremos que determinar a forma irredutível a partir de uma função racional p/q que verifique a condição (4.1). Este procedimento justifica-se pelo facto de que pode acontecer que não existam funções racionais irredutíveis que verifiquem (4.1). Mais exatamente, tem-se

$$(fq - p)(x) = \sum_{i=n+m+1}^{\infty} d_i x^i \;\Rightarrow\; (f - [n/m]_f)(x) = \sum_{i=\mathrm{gr}(p)+\mathrm{gr}(q)+k+1}^{\infty} e_i x^i,$$

onde $\mathrm{gr}(p)$ e $\mathrm{gr}(q)$ representam os graus exatos dos polinómios do numerador e do denominador de $[n/m]_f$ e $k \geq 0$. Estas propriedades, que são facilmente demonstráveis para aproximantes de Padé a uma variável [BGM96], não se verificam geralmente nas várias abordagens da aproximação de Padé em variáveis múltiplas. Por simplicidade apenas se apresentam as várias abordagens de aproximações bidimensionais. As extensões a aproximações de dimensões superiores obtêm-se de forma natural.

Considere-se uma função f em duas variáveis representada pela série de potências

$$f(x, y) = \sum_{(i,j)\in\mathbb{N}_0^2} c_{i,j} x^i y^j. \tag{4.3}$$

Quando passamos a dimensões superiores perde-se a ordem natural existente em $\mathbb{N}_0$ e existem várias formas de agrupar os coeficientes $c_{i,j}$ da série (4.3). Dependendo como agrupamos os coeficientes, c_{ij} obtêm-se diferentes abordagens dos AP bidimensionais. Iremos descrever três abordagens, consideradas por A. Cuyt em [Cuy99]:

- escrevendo (4.3) da forma

$$f(x,y) = \sum_{k=0}^{\infty} c_{i_k,j_k} x^{i_k} y^{j_k}$$

obtêm-se os AP provenientes de *equações definidas por reticulados*;

- reescrevendo (4.3) da forma

$$f(x,y) = \sum_{k=0}^{\infty} \left(\sum_{i+j=k} c_{ij} x^i y^j \right)$$

têm-se os aproximantes *homogéneos*;

- considerando o desenvolvimento em séries de potências na forma

$$f(x,y) = \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\infty} c_{ij} y^j \right) x^i = \sum_{i=0}^{\infty} c_i(y) x^i$$

obtêm-se os chamados AP *nested*;

Iremos de seguida descrever as três abordagens de AP bidimensionais referidas.

4.1.1 AP provenientes de equações definidas por reticulados

Começamos com a seguinte

Definição 4.1.1. Seja A um subconjunto não vazio de $\mathbb{N}_0^2$. Dizemos que A satisfaz a condição de *inclusão* se para todo $(i,j) \in A$, então o conjunto $\{(k,\ell) \mid k \leq i,\ \ell \leq j\} \subset A$.

Seja f uma função a duas variáveis representada pela série de potências

$$f(x,y) = \sum_{(i,j)\in\mathbb{N}_0^2} c_{ij} x^i y^j.$$

Um AP proveniente de equações definidas por reticulados, de f, depende da escolha de três subconjuntos não vazios de índices, N, D e E de $\mathbb{N}_0^2$, os quais satisfazem as seguintes condições

$$N \subset E, \tag{4.4a}$$

$$\operatorname{card}(E \setminus N) = \operatorname{card}(D) - 1, \tag{4.4b}$$

$$E \text{ satisfaz a } \textit{condição de inclusão.} \tag{4.4c}$$

Uma vez escolhidos os conjuntos de índices N, D e E dizemos que a função racional $R_{N,D,E} = p/q$ com

$$p(x,y) = \sum_{(i,j)\in N} a_{i,j} x^i y^j \tag{4.5}$$

$$q(x,y) = \sum_{(i,j)\in D} b_{i,j} x^i y^j \tag{4.6}$$

é um AP proveniente de equações definidas por reticulados, de f, se satisfizer a condição

$$(fq - p)(x,y) = \sum_{(i,j)\in \mathbb{N}_0^2 \setminus E} d_{ij} x^i y^j. \tag{4.7}$$

Deste modo, os coeficientes $a_{i,j}$ e $b_{i,j}$ podem determinar-se resolvendo o sistema de equações lineares homogéneo

$$f(x,y) \sum_{(i,j)\in D} b_{i,j} x^i y^j - \sum_{(i,j)\in N} a_{i,j} x^i y^j = 0, \quad (i,j) \in E. \tag{4.8}$$

A condição (4.4a) permite dividir o sistema de equações lineares (4.8) em dois sistemas,

$$\sum_{k=0}^{i} \sum_{\ell=0}^{j} c_{k\ell} b_{i-k,j-\ell} = a_{ij}, \quad (i,j) \in N \quad \text{(parte não homogénea)} \tag{4.9a}$$

$$\sum_{k=0}^{i} \sum_{\ell=0}^{j} c_{k\ell} b_{i-k,j-\ell} = 0, \quad (i,j) \in E \setminus N \quad \text{(parte homogénea)} \tag{4.9b}$$

onde se considera que $b_{k,\ell} = 0, \quad \text{se } (k,\ell) \notin D$.

A condição (4.4b) garante a existência de um polinómio q não nulo dado que a parte homogénea tem mais uma equação do que incógnitas.

Se a condição de inclusão não se verificar (i.e. existem lacunas no conjunto de índices E) tem-se que a condição (4.7) não implica

$$\left(\frac{1}{q}(fq - p)\right)(x,y) = \left(f - \frac{p}{q}\right)(x,y) = \sum_{(i,j)\in \mathbb{N}_0^2 \setminus E} e_{ij} x^i y^j$$

porque $f - p/q$ contém termos resultantes do produto de termos com índices provenientes das lacunas de E por $(1/q)(x,y)$.

O conjunto das funções racionais p/q que satisfazem (4.7) denomina-se *aproximante de Padé geral* de f e representa-se por $[N/D]_E^f$.

Note-se que no caso unidimensional os AP resultam de se tomar os conjuntos $N = \{0,1,\ldots,n\}$, $D = \{0,1,\ldots,m\}$ e $E = \{0,1,\ldots,n+m\}$ e que esta escolha é única. Quando se faz a extensão a aproximantes com múltiplas variáveis tem-se uma grande variedade de escolhas para os conjuntos de índices N, D e E, o que resulta num grande número de esquemas resultantes desta abordagem.

Propriedades de unicidade e de consistência: Contrariamente ao caso unidimensional esta classe de aproximantes de Padé generalizados não possui geralmente a propriedade de unicidade nem a propriedade de consistência. De facto, podem existir duas funções racionais p_1/q_1 e p_2/q_2 em $[N/D]_E^f$ com formas irredutíveis diferentes. Contudo se as equações homogéneas (4.9b), provenientes do conjunto de índices $D \setminus E$ forem linearmente independentes, pode-se garantir que duas funções racionais p_1/q_1 e p_2/q_2 em $[N/D]_E^f$ têm necessariamente a mesma forma irredutível. Geralmente, no caso multidimensional, apenas podemos afirmar que duas funções racionais p_1/q_1 e p_2/q_2 em $[N/D]_E^f$ satisfazem a relação

$$(p_1q_2 - p_2q_1)(x,y) = \sum_{(i,j)\in(N+D)\setminus E} e_{ij}x^iy^j,$$

onde $N + D = \{(i+k, j+\ell) \mid (i,j) \in N,\ (k,\ell) \in D)\}$, [Cuy99].

A propriedade de consistência está relacionada com a propriedade de unicidade. Neste contexto significa que para uma função racional irredutível

$$f(x,y) = \frac{g(x,y)}{h(x,y)} = \frac{\sum_{(i,j)\in N} g_{ij}x^iy^j}{\sum_{(i,j)\in D} h_{ij}x^iy^j}$$

e para toda a solução $p/q \in [N_k/D_\ell]_{E_{k\ell}}^f$ com $N \subset N_k$ e $D \subset D_\ell$ então p/q e g/h são equivalentes ou seja, tem-se $(ph - gq)(x,y) = 0$. No caso de se verificar a propriedade de consistência então verifica-se a propriedade de unicidade.

4.1.2 AP homogéneos

Dados dois inteiros não negativos m, n considere-se os polinómios

$$A_\ell(x,y) = \sum_{i+j=mn+\ell} a_{ij}x^iy^j, \quad \ell = 0,1,\ldots,m$$

$$B_\ell(x,y) = \sum_{i+j=mn+\ell} b_{ij}x^iy^j, \quad \ell = 0,1,\ldots,n$$

$$C_\ell(x,y) = \sum_{i+j=\ell} c_{ij}x^iy^j, \quad \ell = 0,1,2\ldots$$

Nesta classe de aproximantes de Padé o polinómio $p(x,y)$ do numerador e o polinómio $q(x,y)$ do denominador têm a forma

$$p(x,y) = \sum_{\ell=0}^{m} A_\ell(x,y),$$

$$q(x,y) = \sum_{\ell=0}^{n} B_\ell(x,y),$$

e são calculados usando a condição

$$(fq-p)(x,y)=\sum_{i+j\geq mn+m+n+1} d_{ij}x^iy^j. \tag{4.10}$$

Ou seja, considerando que $C_\ell(x,y)=0$ para $\ell<0$, podemos escrever as equações da condição (4.10) na forma

$$\begin{aligned} C_0(x,y)B_0(x,y)&=A_0(x,y)\\ C_1(x,y)B_0(x,y)+C_0(x,y)B_1(x,y)&=A_1(x,y)\\ &\vdots\\ C_m(x,y)B_0(x,y)+\cdots+C_{m-n}(x,y)B_n(x,y)&=A_m(x,y) \end{aligned} \tag{4.11a}$$

$$\begin{aligned} C_{m+1}(x,y)B_0(x,y)+\cdots+C_{m-n+1}(x,y)B_n(x,y)&=0\\ &\vdots\\ C_{m+n}(x,y)B_0(x,y)+\cdots+C_m(x,y)B_n(x,y)&=0 \end{aligned} \tag{4.11b}$$

Observação: *Neste caso, para funções a duas variáveis, tem-se um sistema com mais uma incógnita do que equações (como no caso unidimensional) no caso geral o sistema é sobre-determinado, mas é possível demonstrar que existe uma solução não trivial. Além disso, os aproximantes homogéneos verificam as propriedades de unicidade e de consistência [Cuy84].*

4.1.3 AP *Nested*

Os AP *Nested* que iremos descrever foram introduzidos por P. Guillaume [Gui97, Gui98]. A principal vantagem destes AP, do ponto de vista da sua implementação, é que o algoritmo apenas usa AP unidimensionais o que se traduz em que se resolvem sistemas de equações lineares de dimensões reduzidas relativamente aos AP acima descritos. Esta vantagem é extensível aos AP *Nested* de séries ortogonais que iremos descrever na próxima secção.

Dada uma função a duas variáveis f na forma $f(x,y)=\sum_{(i,j)\in\mathbb{N}_0^2} c_{ij}x^iy^j$, pode-se tratar a função f como uma função a uma variável considerando a outra variável como sendo um parâmetro. Reescrevendo o desenvolvimento em série de potências da função f na forma

$$\sum_{i=0}^{\infty} c_i(y)x^i \tag{4.12}$$

onde $c_i(y)=\sum_{j=0}^{\infty} c_{ij}y^j$. Pode-se calcular um aproximante de Padé unidimensional de

(4.12). Representando por $[n/m]_x^f$ a forma irredutível de p_x/q_x onde

$$p_x(x,y) = \sum_{i=0}^{n} a_i(y)x^i, \tag{4.13}$$

$$q_x(x,y) = \sum_{i=0}^{m} b_i(y)x^i, \tag{4.14}$$

$$(fq_x - p_x)(x,y) = \sum_{i \geq n+m+1} d_i(y)x^i. \tag{4.15}$$

Se desenvolvermos $[n/m]_x^f$ numa série de potências na variável y

$$[n/m]_x^f(x,y) = \sum_{i=0}^{\infty} \gamma_i(x)y^i, \tag{4.16}$$

onde as funções $\gamma_i(x)$ são funções racionais na variável x, calcula-se um aproximante de Padé unidimensional para (4.16) e representa-se a forma irredutível de p_y/q_y por $[\tilde{n}/\tilde{m}]_y^f \circ [n/m]_x^f$ onde

$$p_y(x,y) = \textstyle\sum_{i=0}^{\tilde{n}} \tilde{a}_i(x)y^i,$$

$$q_y(x,y) = \textstyle\sum_{i=0}^{\tilde{m}} \tilde{b}_i(x)y^i, \tag{4.17}$$

$$\left([n/m]_x^f q_y - p_y\right)(x,y) = \textstyle\sum_{i \geq n+m+1} \tilde{d}_i(x)y^i.$$

Deste modo o cálculo dos AP *nested* bidimensionais $[\tilde{n}/\tilde{m}]_y^f \circ [n/m]_x^f$ exige apenas o cálculo de dois AP $[n/m]_x^f$ e $[\tilde{n}/\tilde{m}]_y^f$ unidimensionais.

Note-se que é possível trocar os papeis das variáveis x e y. Considerando, num primeiro passo, a variável x como parâmetro determinava-se de forma análoga o AP $[n/m]_x^f \circ [\tilde{n}/\tilde{m}]_y^f$. Contudo existe o inconveniente de esta abordagem não tratar as variáveis de f de um modo simétrico. De facto tem-se

$$[\tilde{n}/\tilde{m}]_y^f \circ [n/m]_x^f \neq [n/m]_x^f \circ [\tilde{n}/\tilde{m}]_y^f .$$

Além disso, o numerador e o denominador do AP $[\tilde{n}/\tilde{m}]_y^f \circ [n/m]_x^f$ têm, respetivamente, graus $\tilde{n}$ e $\tilde{m}$ na variável y mas não têm, respetivamente, graus n e m na variável x. Consequentemente estes aproximantes não verificam a propriedade de consistência. Do ponto de vista da abordagem dos AP provenientes de reticulados os polinómios p_y e q_y satisfazem a relação

$$(fq_y - p_y)(x,y) = \sum_{(i,j) \in \mathbb{N}_0^2 \setminus E} d_{i,j}x^i y^j$$

onde o conjunto de índices E contém o conjunto $\{0,1,\ldots,n+m\} \times \{0,1,\ldots,\tilde{n}+\tilde{m}\}$.

As abordagens de AP provenientes de reticulados e de AP *nested* de séries de potências podem generalizar-se de uma forma natural a AP de séries ortogonais. Nas próximas secções iremos descrever os AP bidimensionais de séries ortogonais introduzidas em [Mat07].

4.2 AP bidimensionais de séries ortogonais provenientes de reticulados

Seja f uma função em duas variáveis representada por uma expansão em séries ortogonais

$$f(x,y) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} c_{ij} \phi_i(x) \phi_j(y)$$

onde

- $\{\phi_k\}_{k\geq 0}$ é uma base de polinómios ortogonais no intervalo $[a,b]$ relativamente a uma função peso w
- $c_{i,j} = \frac{1}{\gamma_{i,j}} \int_a^b \int_a^b f(x,y)\phi_i(x)\phi_j(y)w(x)w(y)\mathrm{d}x\mathrm{d}y$
- $\gamma_{i,j} = \|\phi_i\|_w \|\phi_j\|_w$

De forma análoga à AP bidimensional de séries de potências, os AP de séries ortogonais dependem da escolha de três conjuntos de índices N, D e E, os quais devem igualmente satisfazer as condições (4.4a)-(4.4c).

Uma vez escolhidos os conjuntos N, D e E define-se um AP da função f como sendo uma função racional

$$R_{N,D,E}(x,y) = \frac{P_N(x,y)}{Q_D(x,y)} = \frac{\sum\limits_{(i,j)\in N} a_{i,j}\phi_i(x)\phi_j(y)}{\sum\limits_{(i,j)\in D} b_{i,j}\phi_i(x)\phi_j(y)}$$

que satisfaz a condição

$$f(x,y)Q_D(x,y) - P_N(x,y) = \sum_{(i,j)\in \mathbb{N}_0^2 \setminus E} d_{i,j}\phi_i(x)\phi_j(y). \tag{4.18}$$

Note-se que os conjuntos N e D correspondem, respetivamente, aos conjuntos de índices existentes no polinómio do numerador e do denominador do AP e o conjunto E corresponde aos índices para os quais os coeficientes das expansões ortogonais de f e do AP coincidem.

Por uma questão de simplicidade iremos ainda considerar que estes conjuntos satisfazem a seguinte condição suplementar

$$E = N \cup D.$$

Ordenando os conjuntos,

$$
\begin{aligned}
D &= \{(k_m, \ell_m)\}_{m=1,\ldots,\mathrm{card}(D)}\,, \\
N &= \{(i_m, j_m)\}_{m=1,\ldots,\mathrm{card}(N)}\,, \\
H &= E \setminus N = \left\{\left(i_m^{'}, j_m^{'}\right)\right\}_{m=\mathrm{card}(N)+1\ldots,\mathrm{card}(N)+\mathrm{card}(D)-1,}
\end{aligned}
$$

e fazendo

$$
\phi_k(x)\phi_l(y)f(x,y) = \sum_{(i,j)\in\mathbb{N}_0^2} h_{ij}^{kl}\phi_i(x)\phi_j(y), \quad \forall (k,l) \in \mathbb{N}_0^2
$$

onde os coeficientes h_{ij}^{kl} podem calcular-se por recorrência [Mat07]. Dividem-se as equações (4.18) na forma

$$
\sum_{(k,\ell)\in D} h_{ij}^{k\ell} b_{k\ell} = 0, \quad (i,j) \in E \setminus N, \tag{4.19a}
$$

$$
\sum_{(k,\ell)\in D} h_{ij}^{k\ell} b_{k\ell} = a_{ij}, \quad (i,j) \in N. \tag{4.19b}
$$

Como $\mathrm{card}(D) = \mathrm{card}(H)+1$, o sistema (4.19a) é um sistema com $\mathrm{card}(D)-1$ equações lineares em $\mathrm{card}(D)$ incógnitas. Logo existe sempre uma solução não nula para os coeficientes do denominador $(b_{k\ell})_{(k,\ell)\in D}$. Encontrados os coeficientes do denominador encontramos diretamente os coeficientes do numerador $(a_{ij})_{(ij)\in N}$ usando as equações (4.19b).

Supondo que $\mathrm{card}\, N = n$ e $\mathrm{card}\, D = d$ e se ordenarmos os conjuntos dos índices da seguinte forma:

$$
\begin{aligned}
N &= \{(i_1, j_1), \ldots, (i_n, j_n)\}\,, \\
D &= \{(k_1, \ell_1), \ldots, (k_d, \ell_d)\}\,, \\
E &= N \bigcup \{(u_1, v_1), \ldots, (u_{d-1}, v_{d-1})\}
\end{aligned}
$$

o sistema (4.19a), que permite encontrar os coeficientes do denominador, toma a forma

$$
\begin{bmatrix} h_{u_1 v_1}^{k_1 \ell_1} & \cdots & h_{u_1 v_1}^{k_d \ell_d} \\ \vdots & & \vdots \\ h_{u_{d-1} v_{d-1}}^{k_1 \ell_1} & \cdots & h_{u_{d-1} v_{d-1}}^{k_d \ell_d} \end{bmatrix} \begin{bmatrix} b_{k_1 \ell_1} \\ \vdots \\ b_{k_d \ell_d} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.20}
$$

As condições impostas aos conjuntos N, D e E não são restritivas relativamente à sua forma. Deste modo teremos várias classes de aproximantes, consoante a forma escolhida para os conjuntos N, D e E. Destacamos três classes de AP bidimensionais com equações provenientes de reticulados que preservam a simetria entre as duas variáveis [Mat07]: os *aproximantes do tipo produto tensorial quadrados*, os *aproximantes do tipo produto tensorial mistos* e os *aproximantes "homogéneos"*.

4.2.1 AP do tipo tensoriais quadrados

Dados dois inteiros positivos m e ℓ tais que $m \geq \ell$, os aproximantes do tipo tensorial resultam da seguinte escolha para os conjuntos de índices N, D e E:

$$E = \{(i,j)|\ 0 \leq i \leq m-1,\ 0 \leq j \leq m-1\};$$

$$N = \{(i,j)|\ m-\ell \leq i \leq m-1,\ 0 \leq j \leq m-1\};$$

$$\begin{aligned} D = \big\{(i,j)|\ \big(m-\ell < i \leq m-1\ \wedge 0 \leq j \leq m-1\big) \vee \\ \vee \big(0 \leq i \leq m-1\ \wedge m-\ell < j \leq m-1\big)\big\}. \end{aligned}$$

Deste modo os aproximantes têm a seguinte forma,

$$\mathcal{T}_{m,\ell}(x,y) = \frac{\displaystyle\sum_{i=0}^{m-\ell}\sum_{j=0}^{m-\ell} a_{ij}\phi_i(x)\phi_j(y)}{\displaystyle\sum_{i=0}^{m-1}\sum_{j=m-\ell}^{m-1} b_{ij}\phi_i(x)\phi_j(y) + \sum_{i=m-\ell}^{m-1}\sum_{j=0}^{m-1} b_{ij}\phi_i(x)\phi_j(y)}.$$

Trocando os papeis dos conjuntos N e D obtém-se igualmente um aproximante do tipo tensorial.

Para os aproximantes do tipo tensorial podemos definir as seguintes sucessões.

Sucessões verticais Dado um inteiro m positivo, consideramos, neste caso, as seguintes sucessões de conjuntos de índices

$$\begin{aligned} &D = \{(i,j)|\ 0 \leq i,j \leq m\};\\ &N_n = \big\{(i,j)|\big(m+1 \leq i \leq n \wedge 0 \leq j \leq n\big) \vee \big(0 \leq i \leq m \wedge m+1 \leq j \leq n\big)\big\} \bigcup (0,0);\\ &E_n = D \bigcup N_n, \end{aligned}$$

e tem-se a sucessão de aproximantes

$$\mathcal{T}_n^v(x,y) = \frac{P_n(x,y)}{Q(x,y)},\quad n \geq 0,$$

onde os coeficientes do denominador se obtêm resolvendo o sistema (4.20) que devidamente normalizado é um sistema com $m-1$ equações a $m-1$ incógnitas e os coeficientes do numerador são determinados por (4.19b).

Sucessões horizontais Dado um inteiro n fixo consideram-se as seguintes sucessões de conjuntos de índices:

$N = \{(i,j)|\ 0 \leq i,j \leq n\}$;
$D_m = \big\{(i,j)|\big(n+1 \leq i \leq m \wedge 0 \leq j \leq m\big) \vee \big(0 \leq i \leq n \wedge n+1 \leq j \leq m\big)\big\} \bigcup (0,0)$;
$E_m = N \bigcup D_m$.

Temos neste caso aproximantes da forma

$$\mathcal{T}_m^h(x,y) = \frac{P(x,y)}{Q_m(x,y)}, \quad m \geq 0.$$

Ordenando os pares $(i,j) \in \mathbb{N}_0^2$ da seguinte forma:

$$(0,0),(0,1),(1,1),(1,0),(2,0),(2,1),(2,2),(1,2),(0,2),(0,3),\ldots$$

o sistema de equações usado para calcular os coeficientes do denominador de $\mathcal{T}_{m+1}^h$ pode ser obtido à custa do sistema usado para calcular os coeficientes do denominador de $\mathcal{T}_m^h$ acrescentando $2m+1$ linhas e colunas.

Sucessões diagonais Para $m \geq 0$, definimos o aproximante $\mathcal{T}_m^d$ considerando os seguintes conjuntos de índices:

$D_m = \{(i,j)|\ 0 \leq i,j \leq m\}$;
$N_m = \big\{(i,j)|\big(m+1 \leq i \leq 2m \wedge 0 \leq j \leq 2m\big) \vee \big(0 \leq i \leq 2m \wedge m+1 \leq j \leq 2m\big)\big\} \bigcup (0,0)$;
$E_m = \{(i,j)|\ 0 \leq i,j \leq 2m\}$.

Se representarmos os coeficientes do denominador de $\mathcal{T}_m^d$ pela matriz $\mathbf{b_m}$ e representarmos o sistema a resolver na forma matricial por $\mathbf{H_m b_m} = \mathbf{0}$ então os coeficientes do denominador $\mathbf{b_{m+1}}$ de $\mathcal{T}_{m+1}^d$ são determinados pelo sistema $\mathbf{H_{m+1} b_{m+1}} = \mathbf{0}$ com

$$\mathbf{H_{m+1}} = \begin{bmatrix} \mathbf{H_m} & h_{i_2 j_2}^{k_{m^2+1}\ell_{m^2+1}} & \cdots & h_{i_2 j_2}^{k_{m^2+2m+2}\ell_{m^2+2m+2}} \\ & \vdots & & \vdots \\ h_{i_{m^2+1} j_{m^2+1}}^{k_1 \ell_1} & \cdots & \cdots & h_{i_{m^2+1} j_{m^2+1}}^{k_{m^2+2m+2}\ell_{m^2+2m+2}} \\ \vdots & & \vdots & \vdots \\ h_{i_{m^2+2m+2} j_{m^2+2m+2}}^{k_1 \ell_1} & \cdots & \cdots & h_{i_{m^2+2m+2} j_{m^2+2m+2}}^{k_{m^2+2m+2}\ell_{m^2+2m+2}} \end{bmatrix},$$

Então pode-se construir o sistema a partir do anterior acrescentando $2m+1$ linha e colunas.

4.2.2 AP do tipo tensoriais mistos

Optando pelos conjuntos de índices, ver Figura 4.1,

$$\begin{aligned} D_m &= \{(i,j)|\ 0 \leq i,j \leq m\}\,, \\ N_m &= \{(i,j)|\ m+1 \leq i \leq 2m \wedge 0 \leq j \leq m-i\} \cup \\ &\quad \cup \{(i,j)|\ m+1 \leq j \leq 2m \wedge 0 \leq i \leq m-i\} \cup \{(0,0)\}\,, \\ E_m &= N_m \cup D_m \end{aligned}$$

tem-se os AP do tipo tensoriais mistos que iremos representar por $\mathcal{H}_m$.

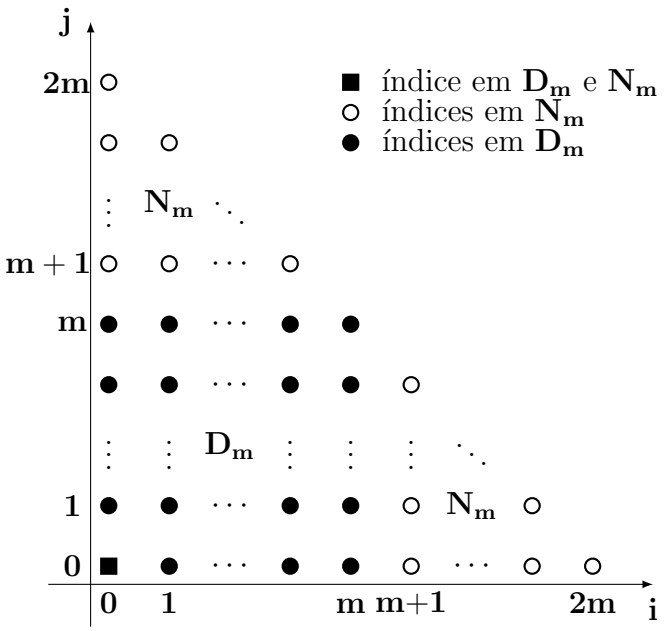


Figura 4.1: Conjuntos de índices para o numerador e denominador dos AP do tipo tensorial mistos.

O cálculo dos AP do tipo tensorial mistos fica especialmente simples para o caso dos polinómios de Chebyshev. Por simplicidade chamaremos a estes aproximantes de *Aproximantes de Chebyshev-Padé mistos* (ACP mistos).

ACP mistos

Na aproximação de Chebyshev-Padé mista consideram-se os polinómios

$$\begin{cases} Q_m(x,y) = \displaystyle\sum_{i=0}^{m}\sum_{j=0}^{m} b_{ij}T_i(x)T_j(y) \\ \\ P_m(x,y) = \displaystyle\sum_{(i,j)\in N_m} a_{ij}T_i(x)T_j(y). \end{cases}$$

Pretende-se encontrar um ACP misto de f, $\mathcal{H}_m = P_m(x,y)/Q_m(x,y)$, de forma a que os polinómios P_m, Q_m com $n \geq m$ satisfaçam a condição

$$Q_m(x,y)f(x,y) - P_n(x,y) = \sum_{(i,j)\in\mathbb{N}_0^2\setminus E_m} r_{ij}T_i(x)T_j(y),$$

onde os conjuntos de índices N_m e D_m estão representados na Figura 4.1.

Usando a seguinte ordenação para conjunto de índices D_m,

$$(0,0),(0,1),\ldots,(0,m),(1,0),(1,1),\ldots,(m,m)$$

as equações (4.19a) formam um sistema homogéneo com m^2 equações a m^2+1 incógnitas. Para se determinar os coeficientes, b_{ij}, $(i,j) \in D_m$, do denominador faz-se $b_{0,0} = 1$ e resolve-se um sistema com m^2 equações a m^2 incógnitas formado pelas equações (4.19b) com a exceção da primeira.

No cálculo dos ACP mistos, em vez de usarmos o algoritmo sugerido em [Mat07], válido para várias famílias de polinómios ortogonais, iremos usar o seguinte algoritmo, especializado para polinómios de Chebyshev e que se fundamenta na seguinte lei de multiplicação de produtos tensoriais de polinómios de Chebyshev bidimensionais

$$\begin{aligned}&(T_i(x)T_j(y))\,(T_r(x)T_s(y)) = \\ &= \frac{1}{4}\left(T_{i+r}(x)T_{j+s}(y) + T_{|i-r|}(x)T_{j+s}(y) + T_{i+r}(x)T_{|j-s|}(y) + T_{|i-r|}(x)T_{|j-s|}(y)\right).\end{aligned} \tag{4.21}$$

Nos pseudocódigos abaixo descritos usaremos, por questões de simplicidade, convenções utilizadas na ferramenta de cálculo MATLAB®.

Pseudocódigo do algoritmo que calcula ACP mistos

1. Dados de entrada: matriz $\mathbf{F}$ cujas entradas, $f_{i,j}$, são os coeficientes da série de Chebyshev de uma função bidimensional; m inteiro positivo que determina os conjuntos N_m, D_m e E_m representados na Figura 4.1
   Dados de saída: Coeficientes do denominador e do numerador guardados, respetivamente, nas matrizes $\mathbf{Q}$ e $\mathbf{P}$

2. Ordenar os elementos (i,j) de D_m, usando a função od definida por

$$\mathrm{od}(i,j) = i(m+1) + j + 1, \quad i,j = 0,1,\ldots,m$$

3. Cálculo dos coeficientes do denominador do ACP misto $\mathcal{H}_m$:

$\mathbf{H} := \text{zeros}((m+1)^2, (m+1)^2)$;
para $i := 0$ **até** m **faça**
  para $j := 0$ **até** m **faça**
    para $r := 0$ **até** $2m$ **faça**
      para $s := 0$ **até** $2m$ **faça**

$$h_{\text{od}(i+r,j+s),\text{od}(i,j)} := \sum_{\substack{i+r\leq m \\ j+s\leq m}} f_{r+1,s+1};$$

$$h_{\text{od}(i+r,|j-s|),\text{od}(i,j)} := \sum_{\substack{i+r\leq m \\ |j-s|\leq m}} f_{r+1,s+1};$$

$$h_{\text{od}(|i-r|,j+s),\text{od}(i,j)} := \sum_{\substack{|i-r|\leq m \\ j+s\leq m}} f_{r+1,s+1};$$

$$h_{\text{od}(|i-r|,|j-s|),\text{od}(i,j)} := \sum_{\substack{|i-r|\leq m \\ |j-s|\leq m}} f_{r+1,s+1};$$

      fim
    fim
  fim
fim
$\mathbf{H}_1 := \mathbf{H}(2 : (m+1)^2, 2 : (m+1)^2)$;
$\mathbf{F}_1 := \mathbf{H}(2 : (m+1)^2, 1)$;
Resolver o sistema de equações lineares $\mathbf{H}_1\mathbf{b}_1 = -\mathbf{F}_1$;
$\mathbf{b} := [1; \mathbf{b}_1]$;
`/* Coeficientes do denominador guardados na matriz coluna` **b** `*/`
para $i := 0$ **até** m **faça**
  $\mathbf{Q}(i+1, :) := \mathbf{b}(\text{od}(i,0) : \text{od}(i,1), 1)^T$
fim
`/* Coeficientes do denominador guardados na matriz quadrada` **Q** `*/`

4. Cálculo dos coeficientes do numerador do ACP misto $\mathcal{H}_m$:

$\mathbf{P} := \text{zeros}(2m+1, 2m+1)$;
$p_{1,1} := \mathbf{H}(1,:)\mathbf{b}$;
para $i := 0$ **até** m **faça**
  para $j := 0$ **até** m **faça**
    para $r := 0$ **até** $3m$ **faça**
      para $s := 0$ **até** $3m$ **faça**
        /* Equações relativas aos índices da parte superior de N_m */
        $p_{i+r+1,j+s+1} := \sum_{\substack{m+1 \leq i+r \leq 2m \\ j+s \leq m-1}} q_{i+1,j+1} f_{r+1,s+1}$;
        $p_{i+r+1,|j-s|+1} := \sum_{\substack{m+1 \leq i+r \leq 2m \\ |j-s| \leq m-1}} q_{i+1,j+1} f_{r+1,s+1}$;
        $p_{|i-r|+1,j+s+1} := \sum_{\substack{m+1 \leq |i-r| \leq 2m \\ j+s \leq m-1}} q_{i+1,j+1} f_{r+1,s+1}$;
        $p_{|i-r|+1,|j-s|+1} := \sum_{\substack{m+1 \leq |i-r| \leq 2m \\ |j-s| \leq m-1}} q_{i+1,j+1} f_{r+1,s+1}$;
        /* Equações relativas aos índices da parte inferior de N_m */
        $p_{i+r+1,j+s+1} := \sum_{\substack{i+r \leq m-1 \\ m+1 \leq j+s \leq 2m}} q_{i+1,j+1} f_{r+1,s+1}$;
        $p_{i+r+1,|j-s|+1} := \sum_{\substack{i+r \leq m-1 \\ m+1 \leq |j-s| \leq 2m}} q_{i+1,j+1} f_{r+1,s+1}$;
        $p_{|i-r|+1,j+s+1} := \sum_{\substack{|i-r| \leq m-1 \\ m+1 \leq j+s \leq 2m}} q_{i+1,j+1} f_{r+1,s+1}$;
        $p_{|i-r|+1,|j-s|+1} := \sum_{\substack{|i-r| \leq m-1 \\ m+1 \leq |j-s| \leq 2m}} q_{i+1,j+1} f_{r+1,s+1}$;
      fim
    fim
  fim
fim
$\mathbf{P} := \frac{1}{4}\mathbf{P}$;
/* Coeficientes do numerador guardados na matriz quadrada $\mathbf{P}$ */

Outra família de AP com equações provenientes de reticulados que podemos construir

é a família dos *AP "homogéneos"*, onde as aspas servem para os distinguir dos AP homogéneos de séries de potências. Embora os polinómios do numerador e do denominador dos AP "homogéneos" não sejam funções homogéneas, como se verifica para os AP homogéneos de séries de potências, adotamos este nome devido à forma dos conjuntos D_m, N_n e $E_{m,n}$ que os definem.

4.2.3 AP "homogéneos"

Os AP "homogéneos" do tipo I, que representaremos por ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$, resultam da escolha dos seguintes conjuntos de índices (ver figura 4.2),

$$\begin{aligned} D_m &= \{(i,j) \mid 0 \leq i+j \leq m\}, \\ N_n &= \{(i,j) \mid m+1 \leq i+j \leq n\} \cup \{(0,0)\}, \\ E_{m,n} &= D_m \cup N_n, \end{aligned}$$

onde m, n são inteiros positivos com $n > m$. Analogamente aos AP tensoriais quadrados, fixando m podem definir-se *sucessões verticais*, fixando $n - m$ definem-se *sucessões horizontais* e considerando $n = 2m$ têm-se as *sucessões diagonais*.

Invertendo os papeis dos conjuntos D_m e N_n obtêm-se os AP "homogéneos" do tipo II, que representamos por ${}^{\mathrm{II}}\mathfrak{H}_{m,n}$.

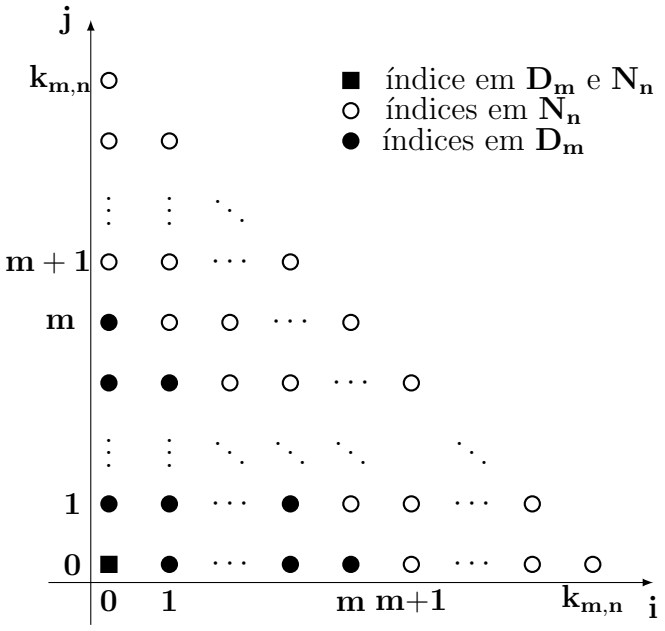


Figura 4.2: Conjuntos de índices para o numerador N_n e para o denominador D_m dos AP "homogéneos" do tipo I, onde $k_{m,n} = m + n + 1$. Invertendo os papeis dos conjuntos N_n e D_m obtêm-se os conjuntos de índices dos AP "homogéneos" do tipo II, em ambos os casos tem-se $E_{m,n} = N_n \cup D_m$.

Logo, normalizando os AP de forma a que o coeficiente de grau 0 do denominador seja unitário, os AP "homogéneos" possuem as seguintes formas,

$$ {}^{\mathrm{I}}\mathfrak{H}_{m,n}(x,y) = \frac{a_{00} + \sum_{m+1\le i+j\le m+n+1} a_{ij}\phi_i(x)\phi_j(y)}{1 + \sum_{1\le i+j\le m} b_{ij}\phi_i(x)\phi_j(y)}, \quad \text{AP "homogéneos" do tipo I,} \quad (4.22) $$

$$ {}^{\mathrm{II}}\mathfrak{H}_{m,n}(x,y) = \frac{\sum_{0\le i+j\le n} a_{ij}\phi_i(x)\phi_j(y)}{1 + \sum_{n+1\le i+j\le m+n+1} b_{ij}\phi_i(x)\phi_j(y)}, \quad \text{AP "homogéneos" do tipo II.} \quad (4.23) $$

De salientar que os AP "homogéneos" do tipo II de uma sucessão vertical partilham todos o mesmo denominador.

Iremos, de seguida, considerar estes aproximantes onde usamos polinómios de Chebyshev na construção dos ACP "homogéneos".

ACP "homogéneos" do tipo I

Usando novamente a lei de multiplicação (4.21), podem determinar-se os ACP "homogéneos" do tipo I, ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$, da forma seguinte: ordena-se o conjunto de índices D_m, indicado na Figura 4.2 e, de modo análogo ao procedimento usado no cálculo dos ACP mistos, usamos as equações (4.19a) para formar o sistema de equações lineares cuja solução determina os coeficientes $b_{i,j}$, $(i,j) \in D_m$, do denominador. Encontrados os coeficientes $b_{i,j}$ determinamos diretamente os coeficientes $a_{i,j}$, $(i,j) \in N_n$ do numerador usando as relações (4.19b). O seguinte pseudocódigo resume o procedimento adotado no cálculo de ACP "homogéneos" do tipo I.

Pseudocódigo para os ACP "homogéneos" do tipo I

1. Dados de entrada: matriz $\mathbf{F}$ cujas entradas, $f_{i,j}$, são os coeficientes da série de Chebyshev de uma função bidimensional; m, n inteiros positivos que determinam os conjuntos N_n, D_m e $E_{m,n}$ representados na Figura 4.2
   Dados de saída: Coeficientes do denominador e do numerador guardados, respetivamente, nas matrizes $\mathbf{Q}$ e $\mathbf{P}$

2. Ordenar os elementos (i,j) de D_m, usando a função od definida por

$$ \mathrm{od}(i,j) = i + 1 + j(m+1) - \frac{(j-1)j}{2}, \quad j = 0,1,\ldots,m, \quad i = 0,1,\ldots,m-j $$

3. Cálculo dos coeficientes do denominador do ACP "homogéneo" do tipo I, ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$:

$\mathbf{H} := zeros(k,k)$, onde $k := (m+1)(m+2)/2$;
para $i = 0$ **até** m **faça**
  para $j = 0$ **até** $m - i$ **faça**
    para $r = 0$ **até** $2m$ **faça**
      para $s = 0$ **até** $2m$ **faça**
        $h_{\mathrm{od}(i+r,j+s),\mathrm{od}(i,j)} := \sum_{i+r+j+s\leq m} f_{r+1,s+1}$;
        $h_{\mathrm{od}(i+r,|j-s|),\mathrm{od}(i,j)} := \sum_{i+r+|j-s|\leq m} f_{r+1,s+1}$;
        $h_{\mathrm{od}(|i-r|,j+s),\mathrm{od}(i,j)} := \sum_{|i-r|+j+s\leq m} f_{r+1,s+1}$;
        $h_{\mathrm{od}(|i-r|,|j-s|),\mathrm{od}(i,j)} := \sum_{|i-r|+|j-s|\leq m} f_{r+1,s+1}$;
      fim
    fim
  fim
fim
$\mathbf{H}_1 := \mathbf{H}(2:k, 2:k)$, $\mathbf{F}_1 := \mathbf{H}(2:k, 1)$;
Resolver o sistema de equações lineares $\mathbf{H}_1\mathbf{b}_1 = -\mathbf{F}_1$;
$\mathbf{b} := [1; \mathbf{b}_1]$;
`/* Coeficientes do denominador guardados na matriz coluna` **b** `*/`
Q := zeros$(m+1, m+1)$;
para $i = 0$ **até** m **faça**
  para $j = 0$ **até** $m - i$ **faça**
    $q_{i+1,j+1} := b_{\mathrm{od}(i,j),1}$;
  fim
fim
`/* Coeficientes do denominador guardados na matriz quadrada` **Q** `*/`

4. Cálculo dos coeficientes do numerador do ACP "homogéneo" do tipo I, ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$:

$\ell := m + n + 1;$
$\mathbf{P} := \text{zeros}(\ell, \ell);$
$p_{1,1} := \sum_{i=1}^{k} b_{1,i} h_{1,i};$
para $i = 0$ **até** m **faça**
  para $j = 0$ **até** $m - i$ **faça**
    para $r = 0$ **até** $n + 2m + 1$ **faça**
      para $s = 0$ **até** $n + 2m + 1$ **faça**

$$p_{i+r+1,j+s+1} := \sum_{m+1 \leq i+r+j+s \leq \ell} q_{i+1,j+1} f_{r+1,s+1};$$

$$p_{i+r+1,|j-s|+1} := \sum_{m+1 \leq i+r+|j-s| \leq \ell} q_{i+1,j+1} f_{r+1,s+1};$$

$$p_{|i-r|+1,j+s+1} := \sum_{m+1 \leq |i-r|+j+s \leq \ell} q_{i+1,j+1} f_{r+1,s+1};$$

$$p_{|i-r|+1,|j-s|+1} := \sum_{m+1 \leq |i-r|+|j-s| \leq \ell} q_{i+1,j+1} f_{r+1,s+1};$$

      fim
    fim
  fim
fim
$\mathbf{P} := \frac{1}{4}\mathbf{P};$
`/* Coeficientes do numerador guardados na matriz quadrada` $\mathbf{P}$ `*/`

ACP "homogéneos" do tipo II

O algoritmo usado no cálculo dos ACP "homogéneos" do tipo II é análogo ao algoritmo, acima descrito usado no cálculo dos ACP do tipo I. Apenas teremos de ter em consideração a troca de papéis dos conjuntos N_n e D_m. Deste modo não mencionaremos, neste trabalho, o pseudocódigo do algoritmo usado para os calcular. No entanto para ajudar a clarificar indicamos na Figura 4.3 os conjuntos de índices N_n e D_m no caso dos AP "homogéneos" do tipo II.

4.3 AP *nested*

Os AP *nested* de séries ortogonais generalizam os AP de séries de potências com o mesmo nome. Iremos aqui referir, os AP *nested* mistos, os quais possuem o polinómio do denominador na base das potências e o polinómio do numerador numa base de polinómios ortogonais, e os ACP *nested* "puros", ou simplesmente, os AP *nested*, os quais possuem os polinómios do numerador e do denominador na base de Chebyshev.

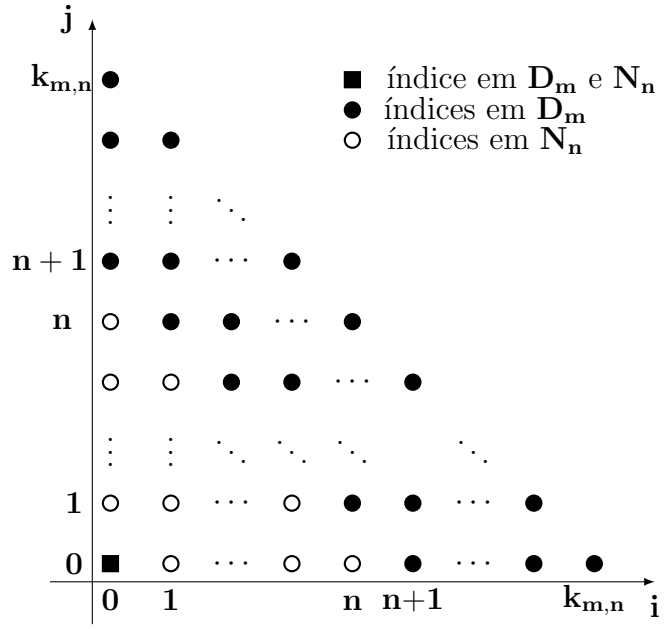


Figura 4.3: Conjuntos de índices para o numerador N_n, denominador D_m dos AP "homogéneos" do tipo II, onde $k_{m,n} = m + n + 1$ e $E_{m,n} = N_n \cup D_m$.

4.3.1 AP *nested* mistos

Seja f uma função representada pela expansão,

$$f(x,y) = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} f_{ij}\phi_i(y)\phi_j(x) = \sum_{j=0}^{\infty} f_j(y)\phi_j(x).$$

O AFP *nested* misto, da função f, é construido em dois passos.

Primeiro passo: Consideramos a função f como uma função na variável x

$$f_y(x) = \sum_{j=0}^{\infty} f_j(y)\phi_j(x).$$

Pretendemos calcular duas funções polinomiais na variável x, Q_y e P_y da forma

$$\begin{cases} Q_y(x) = 1 + \sum_{i=1}^{m} b_i(y)x^i, \\ P_y(x) = \sum_{i=0}^{n} a_i(y)\phi_i(x) \end{cases}$$

que satisfaçam a condição

$$Q_y(x)f_y(x) - P_y(x) = \mathcal{O}\left(\phi_{n+m+1}(x)\right).$$

Substituindo as expressões de $Q_y(x)$ e $P_y(x)$ tem-se

$$\sum_{j=0}^{\infty}\left(f_j(y) + \sum_{i=1}^{m} b_i(y)f_j^i(y)\right)\phi_j(x) - \sum_{i=1}^{n} a_i(y)\phi_i(x) = \mathcal{O}\left(\phi_{n+m+1}(x)\right),$$

onde $f_j^i(y) = x^i f_j(y)$.
O que origina o sistema de equações

$$\sum_{i=1}^{m} b_i(y) f_j^i(y) + f_j(y) = 0, \quad j = n+1, \ldots, n+m \tag{4.24}$$

$$\sum_{i=1}^{m} b_i(y) f_j^i(y) + f_j(y) = a_j(y), \quad j = 0, \ldots, n. \tag{4.25}$$

Segundo passo: Substituímos as funções incógnitas $a_j(y), \quad j = 0, \ldots, n$, $b_i(y), \quad i = 0, \ldots, m$ por polinómios, que as aproximam, da seguinte forma:

I para $i = 1, \ldots, m$ substituímos $b_i(y)$ por um polinómio

$$b_i^*(y) = \sum_{k=0}^{M} b_{ik} y^k.$$

Temos $m(M+1)$ coeficientes a calcular e podemos escolhe-los de modo a anular os primeiros $M+1$ coeficientes da expansão no sistema ortogonal $\{\phi_k\}$ em cada uma das equações em (4.24). Deste modo, os coeficientes $b_{k\ell}$ do denominador são determinados resolvendo o sistema

$$\begin{bmatrix} f_{n+1}^1(y) & f_{n+1}^2(y) & \cdots & f_{n+1}^m(y) \\ f_{n+2}^1(y) & f_{n+2}^2(y) & \cdots & f_{n+2}^m(y) \\ \vdots & \vdots & & \vdots \\ f_{n+m}^1(y) & f_{n+m}^2(y) & \cdots & f_{n+m}^m(y) \end{bmatrix} \begin{bmatrix} b_1^*(y) \\ b_2^*(y) \\ \vdots \\ b_m^*(y) \end{bmatrix} = - \begin{bmatrix} f_{n+1}(y) \\ f_{n+2}(y) \\ \vdots \\ f_{n+m}(y) \end{bmatrix}$$

II Para $j = 0, \ldots, n$ definimos

$$a_j^*(y) = \sum_{k=0}^{N} a_{jk} \phi_k(y)$$

e temos de calcular $(n+1)(N+1)$ coeficientes a_{jk} que serão determinados de forma a satisfazerem as equações (4.25).

Podemos determinar as expansões das funções $f_j^i(y)$ no sistema ortogonal ϕ_k que ocorrem nas equações (4.24) e (4.25), ver [Mat07]. Logo, tem-se que um AP *nested* misto $R(x, y)$ é uma função racional da forma

$$R(x,y) = \frac{P(x,y)}{Q(x,y)} = \frac{\sum_{i=0}^{n} \sum_{k=0}^{N} a_{i,k} \phi_i(x) \phi_j(y)}{1 + \sum_{i=1}^{m} \sum_{k=0}^{M} b_{i,k} x^i y^k}$$

que satisfaz a condição

$$Q(x,y)f(x,y) - P(x,y) = \sum_{(i,j)\in(\mathbb{N}_0^2\setminus E)} e_{ij}\phi_i(x)\phi_j(y)$$

onde

$$E = \{(i,j)|\ (0 \le i \le n\ \wedge\ 0 \le j \le N)\ \vee\ (n+1 \le i \le n+m\ \wedge\ 0 \le j \le M)\}\,.$$

A forma e o cardinal do conjunto de índices E depende dos inteiros n, N, m e M. Se pretendermos simetria relativamente às variáveis x e y escolhe-se $N = M = n + m$ contudo não teremos simetria no aproximante racional $R(x, y)$. Para obtermos simetria em $R(x, y)$ teremos de escolher $N = n$ e $M = m$, e neste caso, privilegiamos a aproximação na variável x.

Iremos de seguida descrever os AP *nested* "puros", para o caso particular em que se usam os polinómios de Chebyshev.

4.3.2 ACP *nested*

Seja f uma função representada pelo desenvolvimento de Chebyshev

$$f(x,y) = \sum_{i,j\ge 0} f_{i,j}T_i(x)T_j(y) = \sum_{j=0}^{\infty} f_j(y)T_j(x), \tag{4.26}$$

e sejam m, n dois inteiros positivos. Para se construir um aproximante de Chebyshev-Padé *nested* $R_{m,n} = P/Q$ da função f, procede-se seguindo os dois passos indicados na secção anterior.

Primeiro passo:

Constroem-se duas funções polinomiais, na variável x, Q_y e P_y da forma

$$P(x,y) = P_y(x) = \sum_{i=0}^{n} a_i(y)T_i(x) \tag{4.27a}$$

$$Q(x,y) = Q_y(x) = \sum_{i=0}^{m} b_i(y)T_i(x) \tag{4.27b}$$

que verificam a condição

$$Q_y(x)f_y(x) - P_y(x) = \mathcal{O}\left(T_{n+m+1}(x)\right).$$

Usando a propriedade dos polinómios de Chebyshev

$$T_i(x)T_j(x) = \frac{1}{2}\left(T_{i+j}(x) + T_{i-j}(x)\right), \quad \forall i,j \geq 0, \text{ onde } T_{-k}(x) = T_k(x), \ \forall k \geq 0$$

tem-se

$$\begin{aligned} Q_y(x)f_j(y) &= \left(\sum_{k=0}^{m} b_k(y)T_k(x)\right)\left(\sum_{i=0}^{\infty} f_i(y)T_i(x)\right) \\ &= \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{k=0}^{m} b_k(y)\big(f_{i-k}(y) + f_{i+k}(y)\big)\right)T_i(x) + \\ &\quad + \frac{1}{2}f_0(y)\sum_{k=0}^{m} b_k(y)T_k(x) + \frac{1}{2}\sum_{k=0}^{m} b_k(y)f_k(y)T_0(x) + \\ &\quad + \frac{1}{2}\sum_{i=n+1}^{\infty}\left(\sum_{k=0}^{m} b_k(y)\big(f_{i-k}(y) + f_{i+k}(y)\big)\right)T_i(x). \end{aligned}$$

Supondo $m \leq n$ e igualando os coeficientes de Chebyshev correspondentes tem-se:

- As funções $b_k(y)$ são soluções do sistema

$$\sum_{k=0}^{m} b_k(y)\left(f_{i-k}(y) + f_{i+k}(y)\right) = 0, \quad i = n+1, \ldots, n+m; \tag{4.28}$$

- As funções $a_k(y)$ satisfazem

$$a_0(y) = \frac{1}{2}\left[b_0(y)f_0(y) + \sum_{k=0}^{m} b_k(y)f_k(y)\right],$$

$$a_i(y) = \frac{1}{2}\left[b_i(y)f_0(y) + \sum_{k=0}^{m} b_k(y)\left(f_{i-k}(y) + f_{i+k}(y)\right)\right], \quad i = 1, \ldots, n.$$

O sistema (4.28) tem m equações e $m+1$ incógnitas. Fixando a função $b_0(y) = 1$ e considerando as séries de Chebyshev-Fourier das funções $b_j(y)$ e $a_k(y)$,

$$b_j(y) = \sum_{i=0}^{\infty} b_{ji}T_i(y), \quad j = 1, \ldots, m, \quad a_k(y) = \sum_{i=0}^{\infty} a_{ki}T_i(y), \quad k = 0, \ldots, n,$$

tem-se que as funções $b_j(y)$ são soluções do sistema com m equações a m incógnitas

$$H^m(y)B(y) = -F(y), \tag{4.29}$$

onde:

$$B(y) = \begin{bmatrix} b_1(y) \\ \vdots \\ b_m(y) \end{bmatrix} = \sum_{i=0}^{\infty} B_i T_i(y) \;\; com \;\; B_i = \begin{bmatrix} b_{1,i} \\ \vdots \\ b_{mi} \end{bmatrix}, \tag{4.30}$$

$$F(y) = 2 \begin{bmatrix} f_{n+1}(y) \\ \vdots \\ f_{n+m}(y) \end{bmatrix}$$

e

$$H^m(y) = \begin{bmatrix} f_n(y) + f_{n+2}(y) & f_{n-1}(y) + f_{n+3}(y) & \cdots & f_{n+1-m}(y) + f_{n+1+m}(y) \\ \vdots & \vdots & & \vdots \\ f_{n+m-1}(y) + f_{n+m+1}(y) & f_{n+m-2}(y) + f_{n+m+2}(y) & \cdots & f_n(y) + f_{n+2m}(y) \end{bmatrix}.$$

Segundo passo:

Substituem-se as séries de Chebyshev das funções $b_i(y)$ e $a_i(y)$ pelas suas aproximações polinomiais $b_i^*(y)$ e $a_i^*(y)$.

Para calcularmos os coeficientes $b_{i,j}$ substituímos B por um polinómio B^* de grau m, ou seja trunca-se a expansão em (4.30) a partir da ordem m,

$$B^*(y) = \begin{bmatrix} b_1^*(y) \\ \vdots \\ b_m^*(y) \end{bmatrix} = \sum_{i=0}^{\infty} B_i^* T_i(y).$$

A expansão na série de Chebyshev do lado esquerdo de (4.29) toma a forma

$$H^m(y)B^*(y) = \frac{1}{2}\sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} H_{k-j}B_j^*\right)T_k(y) + \frac{1}{2}\sum_{k=1}^{m}\left(\sum_{j=k}^{m} H_{k-j}B_j^*\right)T_k(y) + \\ + \frac{1}{2}\sum_{k=0}^{\infty}\left(\sum_{j=0}^{m} H_{k+j}B_j^*\right)T_k(y),$$

onde para todo $i \geq 0$, $H_i \in \mathcal{M}_{m\times m}$ é o i-ésimo coeficiente da série de Chebyshev de $H^m(y)$.

Considerando F_k o coeficiente de índice k da expansão de $F(y)$ e reagrupando os coeficientes com o mesmo índice k na expansão no sistema $\{T_k\}$ e substituindo H_0 e F_0 por

$\frac{1}{2}H_0$ e $\frac{1}{2}F_0$ obtém-se o sistema de equações lineares

$$\begin{cases} 2\sum_{j=0}^{m} H_j B_j^* = -F_0, \\ \sum_{j=0}^{m} H_{j-k} B_j^* + \sum_{j=0}^{m} H_{j+k} B_j^* = -F_k, \quad k = 1, \ldots, m \end{cases}$$

onde se considera $H_{-i} = H_i$.

Teorema 4.3.1 ([Mat07])**.** Seja f uma função a duas variáveis, dada pela expansão em série de Chebyshev (4.26). Então os polinómios P e Q, definidos em (4.27a) e (4.27b), construidos pelos dois passos acima indicados satisfazem a condição

$$P(x,y)f(x,y) - Q(x,y) = \sum_{i,j\geq 0} r_{ij} T_i(x) T_j(y)$$

com $r_{ij} = 0$ para $i = 0, \ldots, n+m$ e $j = 0, \ldots, m$.

Nas próximas secções iremos testar numericamente o comportamentos dos ACP bidimensionais descritos nas secções anteriores. Analogamente aos testes unidimensionais iremos comparar os ACP com séries truncadas e usaremos os ACP bidimensionais para filtrar soluções espectrais da equação de Poisson bidimensional.

4.4 Testes Numéricos de ACP *nested*

Analogamente ao caso unidimensional começamos por estudar o efeito da perturbação nos coeficientes numa série de Chebyshev bidimensional nos ACP *nested*.

4.4.1 ACP *nested* de séries perturbadas

Consideremos a série de Chebyshev bidimensional

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} f_{i,j} T_i(x) T_j(y) \tag{4.31}$$

com os coeficientes definidos por

$$f_{i,j} = \begin{cases} f_{0,0} = \frac{1}{4} & \\ f_{2i+1,0} = \frac{(-1)^i}{(2i+1)\pi} & i > 0 \\ f_{0,2j+1} = \frac{(-1)^j}{(2j+1)\pi} & j > 0 \\ f_{2i+1,2j+1} = (-1)^{i+j} \frac{4}{(2i+1)(2j+1)\pi^2} & i,j > 0 \\ 0 & \text{casos restantes} \end{cases}$$

que converge pontualmente no quadrado $\Omega =]-1,1[^2$ para a função descontínua

$$f(x,y) = \begin{cases} 0 & \text{se } x < 0 \vee y < 0 \\ \frac{1}{2} & \text{se } (x = 0 \wedge y \geq 0) \vee (y = 0 \wedge x \geq 0) \\ 1 & \text{se } x > 0 \wedge y > 0 \end{cases} \tag{4.32}$$

que designaremos por *função salto*.

Tendo em vista analisar a qualidade dos AP iremos comparar os erros absolutos dos ACP *nested* $R_{m,n}$ com os erros absolutos da série truncada

$$f_{m,n}(x,y) = \sum_{i=0}^{2m+n+1} \sum_{j=0}^{m+n+1} f_{i,j} T_i(x) T_j(y), \tag{4.33}$$

definidos por

$$\begin{aligned} \Delta R_{m,n} &= |f - R_{m,n}| \quad \text{(erro absoluto do ACP } \textit{nested}) \\ \Delta f_{m,n} &= |f - f_{m,n}| \quad \text{(erro absoluto da série truncada)} \end{aligned}$$

Note-se, que o cálculo de $R_{m,n}$ usa todos os coeficientes do polinómio $f_{m,n}(x,y)$.

Na Figura 4.4 ilustramos os erros dos AP $R_{3,3}$ e de $f_{3,3}$ em Ω. Podemos observar a assimetria existente nos erros $\Delta R_{3,3}$ e $\Delta f_{3,3}$ devido ao facto dos ACP *nested* exibirem uma aproximação assimétrica. Podemos igualmente observar que na vizinhança dos pontos de descontinuidade da função f, a AP $R_{3,3}$ reduz as oscilações, causadas pelo fenómeno de Gibbs, exibidas por $f_{3,3}$. Esta observação fica mais clara analisando a Figura 4.5 onde representamos os erros $\Delta R_{3,3}$ (curva continua) e $\Delta f_{3,3}$ (curva a tracejado) em 6 secções de Ω que atravessam o salto da função f. Na coluna da esquerda apresentamos 3 secções horizontais para valores de $y = 1/2, -1/2$ e $y = 0$ com $-1 \leq x \leq 1$ e na coluna da direita apresentamos 3 secções verticais para valores de $x = 1/2, -1/2$ e $x = 0$ com $-1 \leq y \leq 1$. As Figuras 4.6 e 4.7 ilustram os gráficos dos erros absolutos $\Delta R_{m,n}$ e $\Delta f_{m,n}$ representados, respetivamente, nas Figuras 4.4 e 4.5) mas para valores de $m = n = 11$.

No estudo de ACP *nested* de séries perturbadas, iremos efetuar uma análise da localização dos pólos e zeros destes aproximantes calculados a partir de perturbações nos coeficientes da série (4.31). Analogamente ao caso unidimensional iremos considerar dois tipos de ruídos

$$T_\epsilon(x,y) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \epsilon r_{i,j} T_i(x) T_j(y), \quad \text{do tipo I} \tag{4.34}$$

$$T_\omega(x,y) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \omega r_{i,j} 2^{-(i+j)} T_i(x) T_j(y), \quad \text{do tipo II} \tag{4.35}$$

onde os coeficientes $r_{i,j}$ são números aleatórios uniformemente distribuídos no intervalo $[-1,1]$ e ϵ, ω são valores positivos que representam a "força" do ruído.

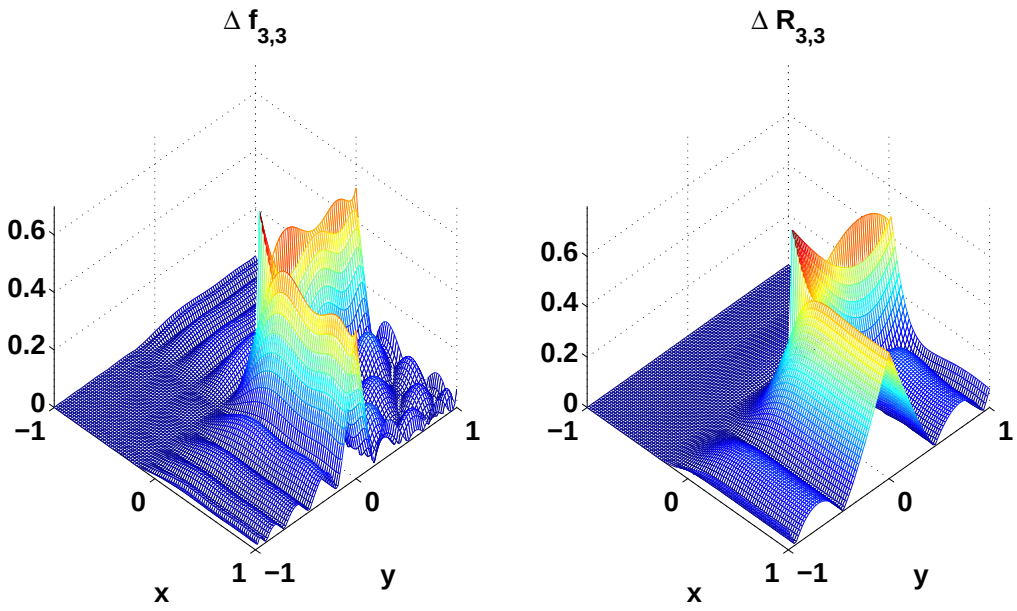


Figura 4.4: Erros absolutos da série de Chebyshev truncada (à esquerda) e da aproximação de Chebyshev Padé *nested* (à direita) com $m = n = 3$.

Na Figura 4.8 ilustramos as curvas de cor preta, que representam os pólos, e as curvas de cor magenta, que representam os zeros, de ACP *nested*, $R_{5,5}$ (imagens da linha superior), $R_{7,7}$ (imagens da linha central) e de $R_{11,11}$ (imagens da linha inferior). Os aproximantes, da coluna à direita são ACP *nested* da série (4.31), isenta de perturbação, os da coluna central são da série (4.31) perturbada com ruído do tipo I e os aproximantes da coluna à esquerda são da mesma série mas perturbada com ruído do tipo II. Numa primeira análise salientamos a seguinte relação entre os ruídos bidimensionais (4.34) e (4.35) e os seus homólogos unidimensionais (2.34) e (2.35) para polinómios de Chebyshev: os ruídos do tipo (2.34) originam pares de Froissart nos AP no intervalo $[-1, 1]$ e o seu homólogo bidimensional origina pares de Froissart nos ACP *nested* em Ω. Relativamente aos ruídos do tipo (2.35) originam pares de Froissart localizados na elipse de Bernstein $\mathcal{E}_{r(2)}$ e os pólos/zeros do ruído homólogo dimensional não estão localizados em Ω. Esta relação deve-se ao facto de a base bidimensional ser construída via produto tensorial dos elementos da base unidimensional. Numa segunda análise notamos que no caso bidimen-

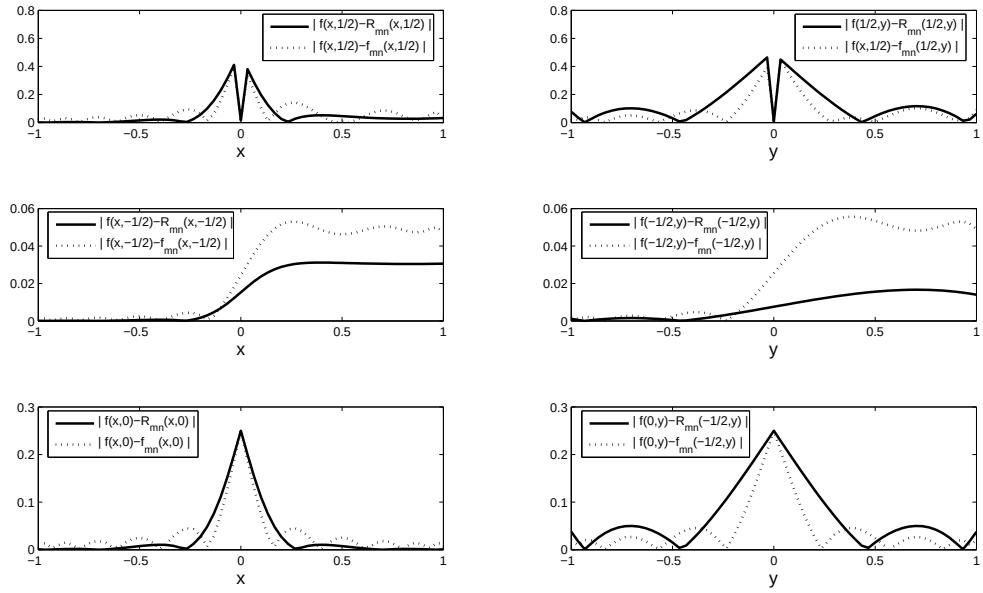


Figura 4.5: Pormenor dos erros da série de Chebyshev truncada (linhas ponteadas) e da aproximação de Chebyshev Padé *nested* (linhas contínuas) das Figuras 4.4. Na coluna à esquerda representa-se os erros nos intervalos $y = 1/2$, $y = -1/2$ e $y = 0$ e na coluna à direita os erros nos intervalos $x = 1/2$, $x = -1/2$ e $x = 0$.

sional os pares de Froissart e os pólos espúrios aparecem para AP de ordens reduzidas e para ruídos com valores de ϵ de ordem inferiores (neste caso usou-se $\epsilon = \omega = 10^{-6}$). Este facto está relacionado com o mau condicionamento e com a dimensão das matrizes envolvidas no cálculo dos ACP *nested*. Estas observações sugerem que a filtragem, usando ACP *nested*, apenas será eficaz se a qualidade dada pelos filtros $R_{m,n}^{(N,M)}$ for superior à qualidade da soluções espectrais $y_{N,M}$, para valores de N e M superiores aos valores de n e m, sendo que os valores n e m são necessariamente pequenos, em virtude da localização dos pólos/zeros.

4.4.2 Filtragem espectral via ACP *nested*

Nos processos de filtragem via ACP bidimensionais, iremos resolver a equação de Poisson bidimensional usando para o efeito o esquema de colocação descrito no exemplo 1.4.2. Consideramos nas equações (1.42a) e (1.42b) as funções g, h_W, h_E, h_S e h_N definidas

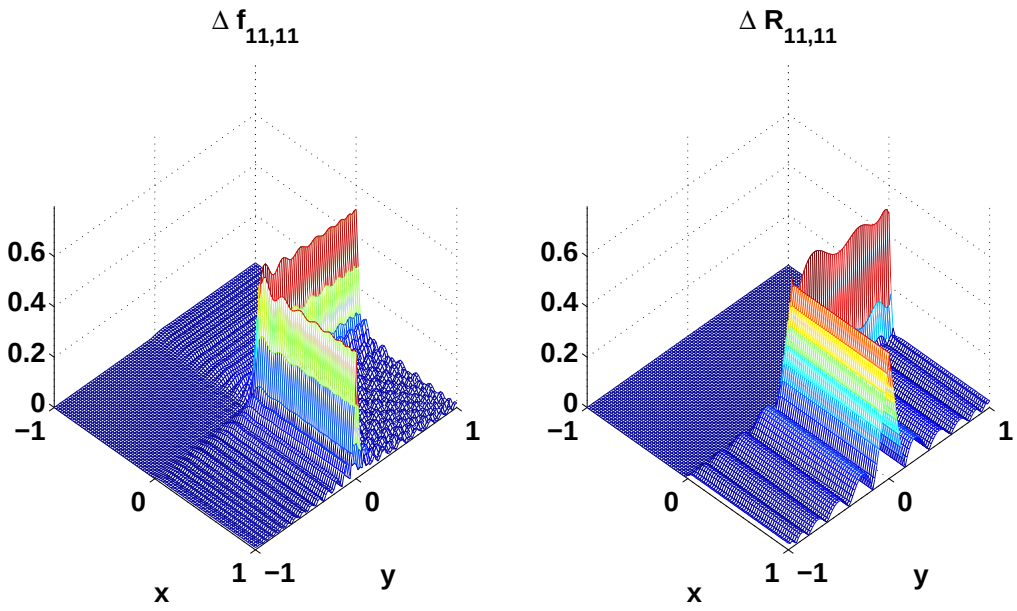


Figura 4.6: Erros absolutos da série de Chebyshev truncada (à direita) e da aproximação de Chebyshev Padé *nested* (à esquerda) com $m = n = 11$.

respetivamente por

$$g(x,y) = ((x-a)^2 + (y-b)^2)^{-1/2} \tag{4.36}$$

$$h_W(y) = ((1+a)^2 + (y-b)^2)^{1/2} \tag{4.37}$$

$$h_E(y) = ((1-a)^2 + (y-b)^2)^{1/2} \tag{4.38}$$

$$h_N(x) = ((1-b)^2 + (x-a)^2)^{1/2} \tag{4.39}$$

$$h_S(x) = ((1+b)^2 + (x-a)^2)^{1/2} \tag{4.40}$$

onde a e b são parâmetros reais. Esta escolha justifica-se pelo facto deste problema ter como solução a função

$$u(x,y) = ((x-a)^2 + (y-b)^2)^{1/2},$$

a qual possui uma singularidade no plano real, em $s = (a,b)$. Deste modo, podemos controlar a localização da singularidade e estudar o comportamento dos vários filtros para diferentes localizações do ponto s.

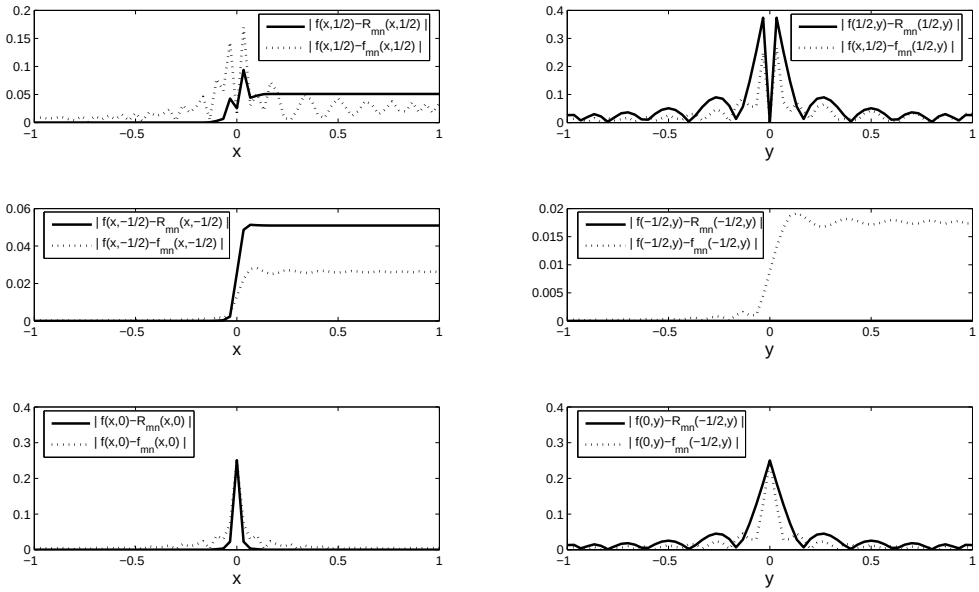


Figura 4.7: Pormenor dos erros da série de Chebyshev truncada e da aproximação de Chebyshev Padé *nested* das Figuras 4.6.

Neste exemplo, optámos por tomar para valores dos parâmetros $a = 0$ e $b = 0$ e chamaremos, ao longo deste capítulo, a esta equação de *problema A* o qual possui uma solução com uma singularidade, $s_{\mathrm{A}} = (0, 0)$, no interior do domínio Ω. Dado que a função u não é derivável no ponto s_{A} a taxa de convergência do método de colocação não será exponencial. Na Figura 4.9 ilustramos os erro absolutos $\Delta u_N = |u - u_N|$ das soluções de colocação de ordem N

$$u_N(x, y) = \sum_{i=0}^{N} \sum_{j=0}^{N} u_{i,j}^{(N)} T_i(x) T_j(y)$$

para valores de $N = 20, 40, 60$. De facto, os erros absolutos Δu_N atingem os seus valores máximos $1.78e{-}1$, $1.01e{-}1$, e $7.29e{-}2$ para valores de $N = 20, 30$ e $N = 60$, respetivamente, em pontos perto da singularidade de u, s_A. Além da taxa de convergência ser lenta, este esquema de colocação, é numericamente instável, perdendo precisão para valores de $N > 60$. Logo, será convenientes usar filtros calculados com os coeficientes de u_{60} para, deste modo, usarmos coeficientes afetados com o mínimo ruído possível.

Em face do que foi dito na secção anterior, iremos analisar a localização dos pólos e zeros de ACP diagonais $R_{n,n}^{(60)}$ para os valores de n disponíveis, ou seja, para valores de $n \leq 19$. A localização dos pólos/zeros de $R_{n,n}^{(60)}$ encontra-se ilustrada na Figura 4.10 para valores de n entre 2 e 10. Podemos observar dois padrões distintos conforme a paridade de n. Para valores de n ímpares existe uma curva de zeros no segmento $x = 0$ o que só

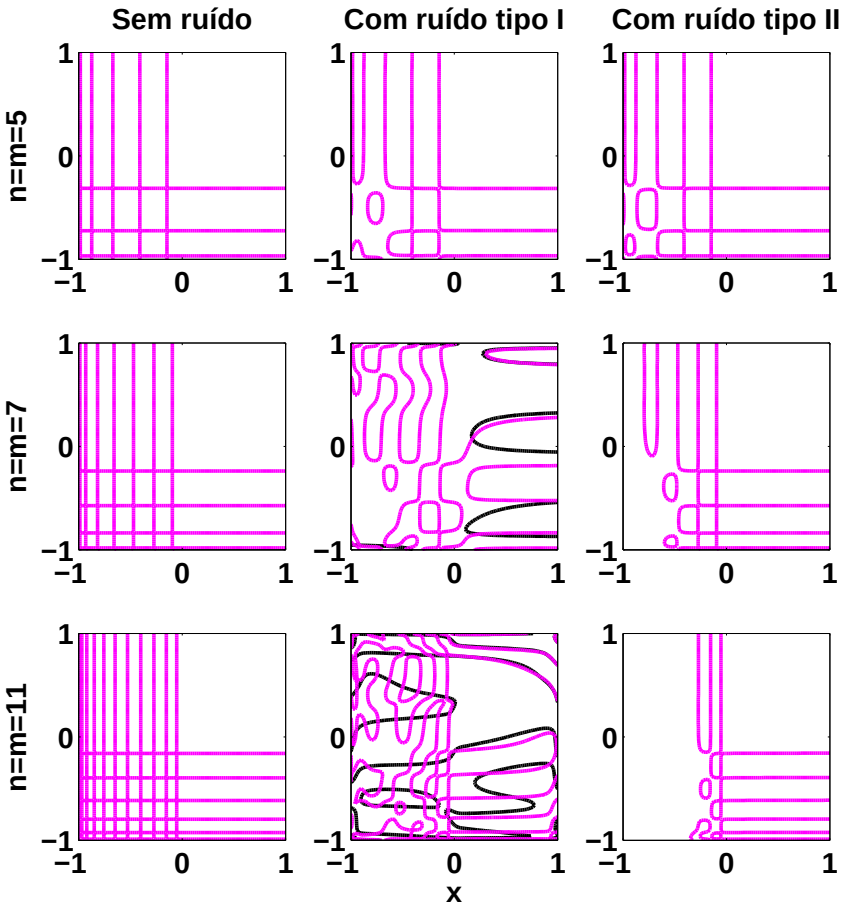


Figura 4.8: Localização dos pólos (curvas a preto) e dos zeros (curvas a magenta) de ACP nested $R_{n,m}$, $n = m = 5, 7$ e 11, da série (4.31) sem perturbação e com perturbações do tipo I e do tipo II, com $\epsilon = \omega = 10^{-6}$.

por si arruína a qualidade da aproximação destes filtros. Note-se que u apenas possui um zero na origem, que coincide com a sua singularidade. Numa segunda análise pode-se observar que existem curvas de pólos que coincidem na figura com curvas de zeros, ou seja, existem pares de Froissart bidimensionais. Estes pares vão proliferando quando o valor de n aumenta, ver Figuras 4.10 e 4.11. A Figura 4.11 é a continuação da Figura 4.10 no sentido que ilustra a localização dos pólos/zeros para os valores N de 11 até 19 e por razões de boa visibilidade entendeu-se separar os resultados em duas figuras. Neste exemplo, não é difícil encontrar o melhor filtro diagonal. Com efeito, basta observar que apenas os aproximantes $R^{(60)}_{(4,4)}$ e $R^{(60)}_{(6,6)}$ não possuem pólos espúrios nem pares de Froissart em Ω, logo $R^{(60)}_{(6,6)}$ será o melhor aproximante diagonal.

Podemos observar na Figura 4.12 que o filtro $R^{(60)}_{6,6}$ é de facto o melhor aproximante

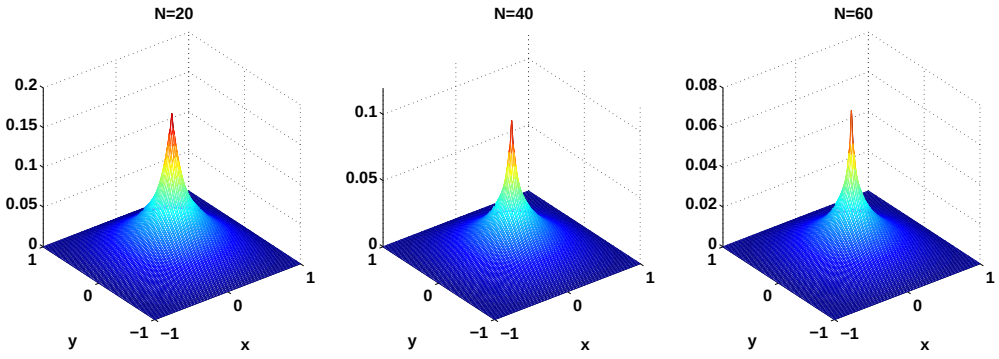


Figura 4.9: Erros absolutos Δu_N das soluções de colocação de ordens $N = 20, 40, 60$ do problema A.

diagonal, contudo a sua aproximação não melhora a aproximação dada pela solução espectral u_{60} cujo erro é apresentado na Figura 4.9. Utilizando o mesmo aproximante de colocação u_{60}, analisámos diversos filtros $R^{(60)}_{m,n}$ não diagonais e não foi possível encontrar um que melhore o resultado obtido por $R^{(60)}_{6,6}$. Este exemplo é paradigmático no que diz respeito aos ACP *nested*, pelo menos, nas experiências efetuadas. Mesmo para problemas com soluções suaves, observaram-se conclusões idênticas. Ou seja, os ACP diagonais são muito sensíveis a pequenos ruídos nos coeficientes espectrais o que provoca a existência de pares de Froissart (e de pólos espúrios) que arruínam a qualidade da aproximação.

Na próxima secção iremos estudar o comportamentos dos filtros com equações provenientes de reticulados.

4.5 Testes Numéricos de ACP com equações provenientes de reticulados

Ao longo desta secção iremos estudar o comportamento dos ACP do tipo tensorial mistos, ou, abreviadamente, *ACP mistos* $\mathcal{H}_m$, dos ACP homogéneos do tipo I ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$ e dos ACP homogéneos do tipo II ${}^{\mathrm{II}}\mathfrak{H}_{m,n}$. Para o efeito, iremos ilustrar os resultados utilizando a função salto (4.32) e quatro problemas derivados da equação de Poisson (1.42a), (1.42b) com a função g definida por (4.36), e as condições fronteira h_W, h_E, h_N e h_S definidas, respetivamente, por (4.37), (4.38), (4.39) e (4.40).

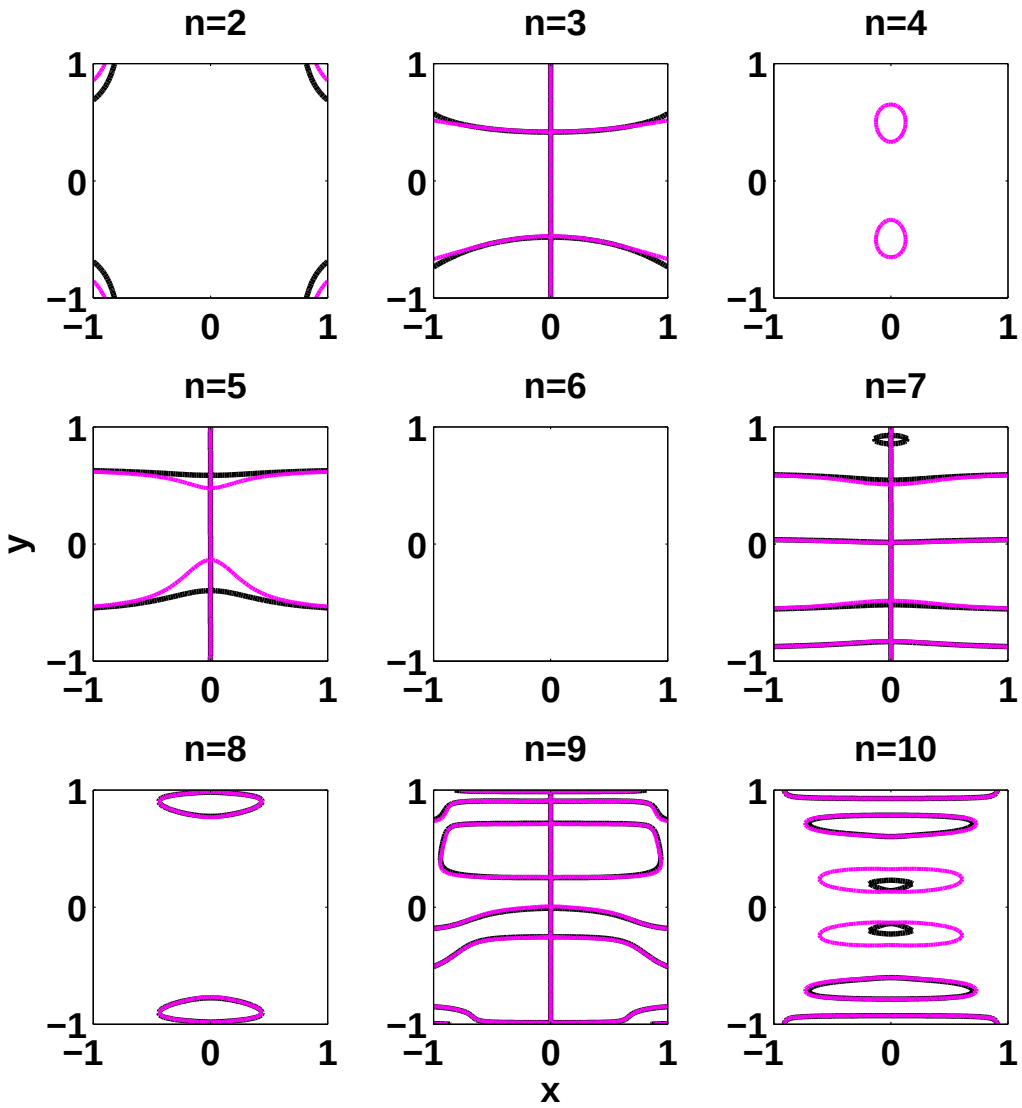


Figura 4.10: Localização dos zeros (curva de cor magenta) e pólos (curva de cor preta) dos AP *nested* $R_{n,n}^{(60)}$ acessíveis, para valores de $n = 2, 3, \ldots, 10$ do problema A. Para os valores de n superiores ver Figura 4.10.

O objetivo é analisar o comportamento de filtros com equações provenientes de reticulados quando movemos a singularidade $s = (a, b)$ da solução deste problema de Poisson, $u(x, y) = \sqrt{(x-a)^2 + (y-b)^2}$, por diferentes regiões do domínio $\Omega =]-1, 1[^2$. Com efeito, iremos considerar os seguintes problemas:

1. problema A, já considerado na secção anterior, com singularidade $s_{\mathrm{A}} = (0, 0)$ localizada na interseção das diagonais do domínio Ω

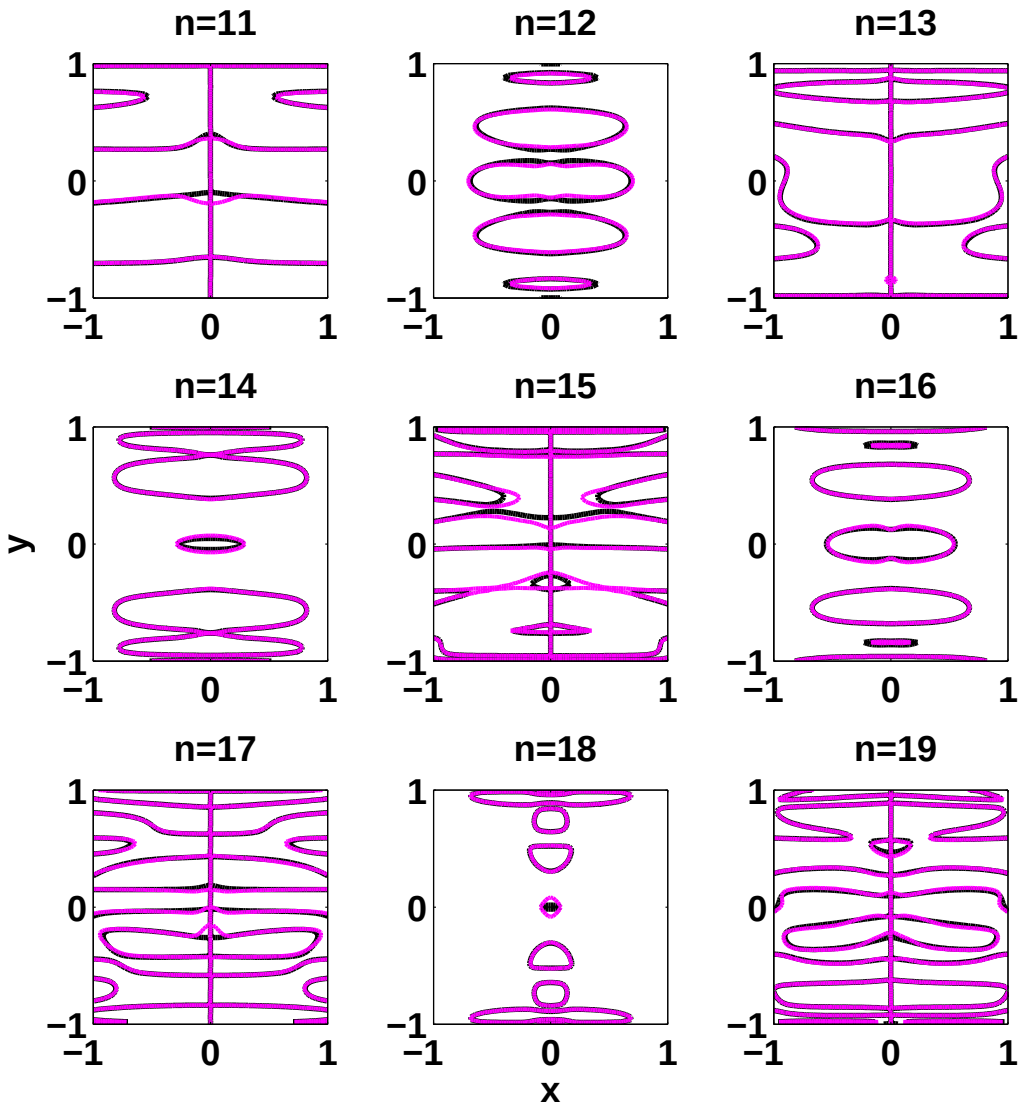


Figura 4.11: Continuação dos resultados apresentados na Figura 4.10 para os restantes valores de n acessíveis.

2. problema B, com singularidade $s_{\mathrm{B}} = (1/2, -3/4)$ localizada no interior de Ω e próxima da fronteira de Ω

3. problema C, com singularidade $s_{\mathrm{C}} = (1/2, 1)$ localizada na fronteira de Ω

4. problema D, com singularidade $s_{\mathrm{D}} = (1, 1)$ localizada num vértice da fronteira de Ω.

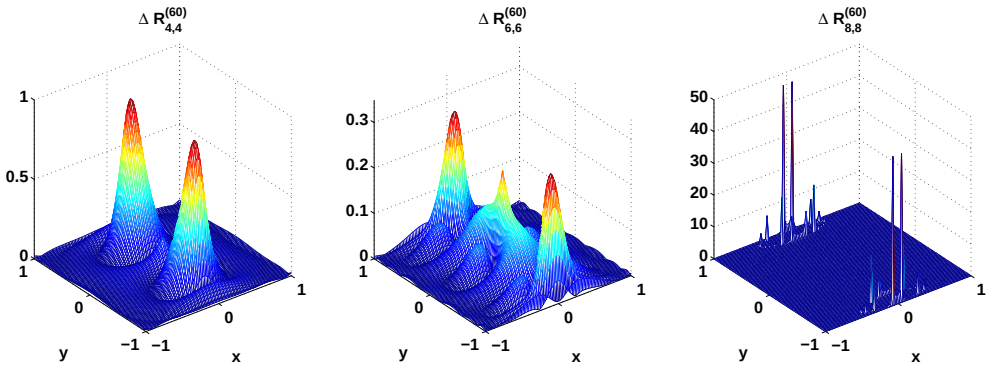


Figura 4.12: Erros absolutos de filtros *nested* diagonais $R_{n,n}^{(60)}$ para valores de $n = 4, 6, 8$ do problema A.

4.5.1 ACP mistos

ACP mistos da função salto

Para testar o comportamento do ACP mistos, considerámos a série (4.31) da função salto e começámos por observar a localização dos pólos e zeros de AP $\mathcal{H}_m$ para diferentes valores de m. Os resultados obtidos não foram promissores dado que, para todos os valores de m testados os aproximantes $\mathcal{H}_m$ possuem pares de Froissart e pólos espúrios. Ilustramos este comportamento na Figura 4.13 na qual apresentamos a localização de pólos e zeros dos AP $\mathcal{H}_m$ para valores de $m = 1, 2 \ldots, 9$ em Ω. Pode-se observar a simetria existente nas curvas que representam os pólos e zeros para valores de $m \leq 5$. Esta observação justifica-se pelo facto de usarmos coeficientes com ruído que podemos considerar negligenciável, por resultarem apenas da resolução do sistema de equações lineares (4.20) que não são mal condicionados para $m \leq 5$. Para valores de $m > 5$ os sistemas de equações lineares (4.20) são mal condicionados e introduzem ruídos nos AP o que provoca a quebra da simetria. Note-se que estes sistemas possuem, após a normalização $b_{0,0} = 1$, $m^2 - 1$ equações em $m^2 - 1$ incógnitas.

> **Observação:** *Os resultados obtidos de ACP provenientes de reticulados com a série* (4.31) *da função salto, perturbada com ruído do tipo* I (4.34) *e do tipo* II (4.35) *foram análogos aos resultados dos ACP nested das séries perturbadas* (4.34) *e* (4.35)*, ilustrados na Figura 4.8 e, por isso, não são aqui apresentados.*

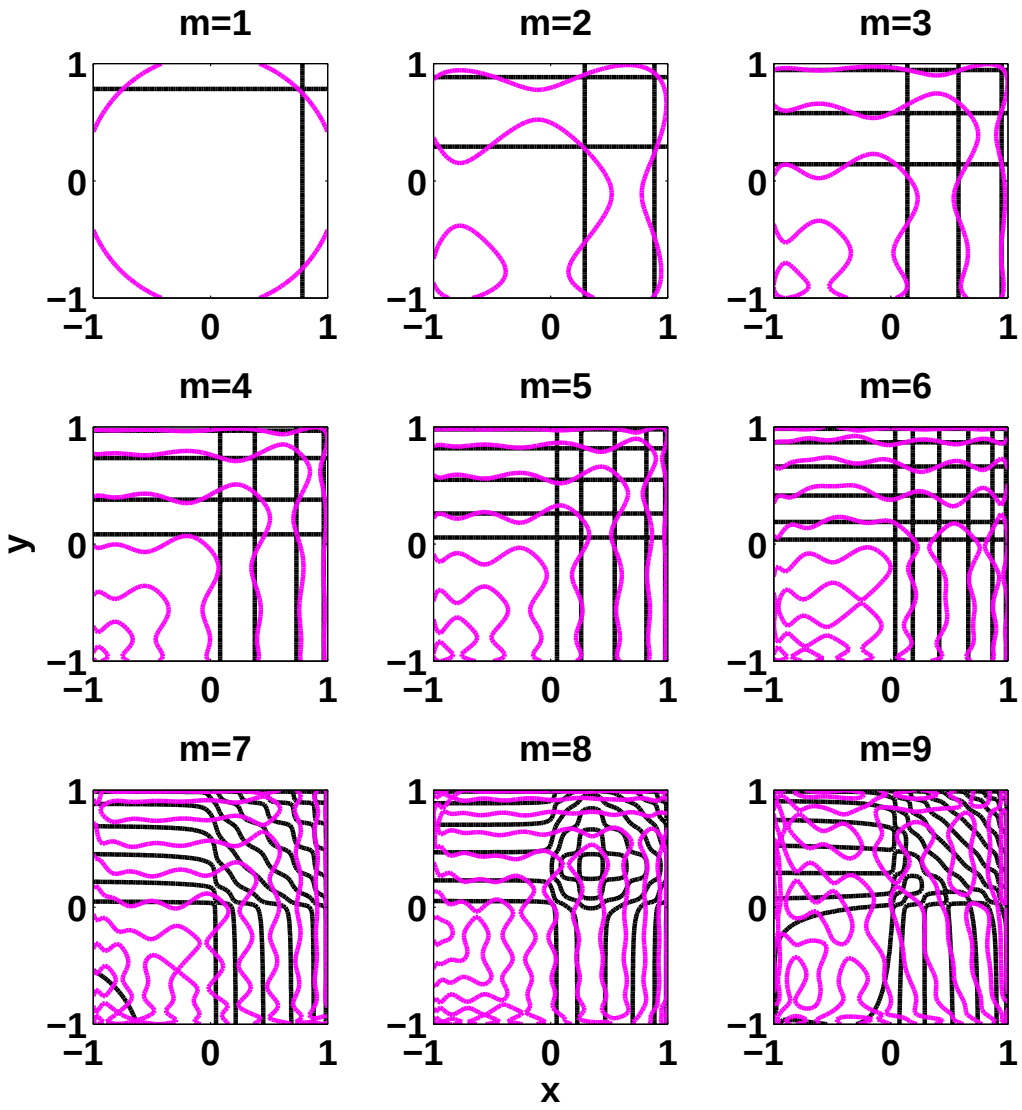


Figura 4.13: Localização de pólos (curvas a preto) e zeros (curvas a magenta) de aproximantes tensoriais mistos $\mathcal{H}_m$, $m = 1, 2, \ldots, 9$, da função salto.

Filtragem via ACP mistos

Tendo em vista testar o comportamento dos ACP mistos na filtragem de uma solução espectral e comparar com os resultados obtidos para os filtros *nested*, começamos por considerar novamente o problema A, com singularidade no ponto $s_{\mathrm{A}} = (0, 0)$. Usamos, para o efeito, os filtros mistos disponíveis, $\mathcal{H}_m^{(60)}$, $m = 1, 2, \ldots, 19$ e os respetivos erros absolutos

$$\Delta\mathcal{H}_m^{(60)} = \left|u - \mathcal{H}_m^{(60)}\right|, \quad m = 1, 2, \ldots, 19.$$

O primeiro procedimento, verificar a localização dos pólos e zeros dos filtros mistos $\mathcal{H}_m^{(60)}$, revelou que nenhum dos filtros disponíveis possuía pólos ou zeros em Ω. Os resultados dos erros $\Delta\mathcal{H}_m^{(60)}$ revelaram que a aproximação fornecida pelos filtros melhora quando se aumenta o valor de m. Contudo os ACP mistos não melhoram a aproximação dada pela solução espectral u_{60}. Este facto encontra-se ilustrado na Figura 4.14 onde apresentamos os gráficos dos erros dos filtros $\mathcal{H}_m^{(60)}$, $m = 1, 10$ e 19. Em termos de erro máximo absoluto verifica-se que o processo de filtragem também não consegue melhorar a aproximação da solução espectral. De facto tem-se $\max\limits_{(x,y)\in\Omega} \Delta u_{60} = 7.29e{-}2$ e $\max\limits_{(x,y)\in\Omega} \Delta\mathcal{H}_{19}^{(60)} = 8.12e{-}2$. Concluímos deste modo que não há vantagens em filtrar soluções de colocação do problema A usando ACP mistos.

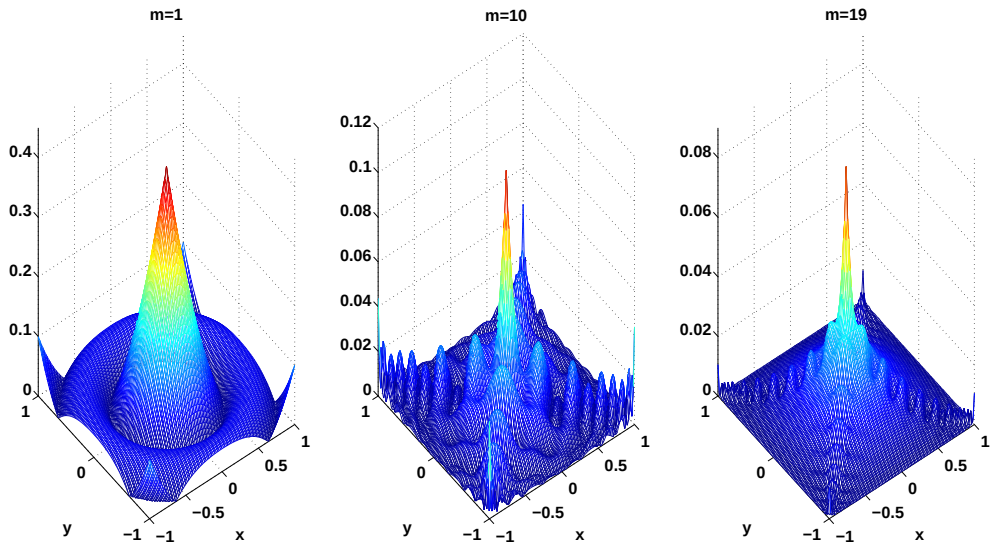


Figura 4.14: Erros absolutos dos filtros mistos $\Delta\mathcal{H}_m^{(60)}$, $m = 1, 10$ e 19, do problema A.

Tendo em vista, estudar o comportamento do método de colocação e dos filtros mistos a soluções com uma singularidade perto da fronteira de Ω iremos agora considerar o problema B, com singularidade no ponto $s_{\mathrm{B}} = (1/2, -3/4)$. De modo análogo ao problema A, também neste problema se observou que o método de colocação é numericamente estável até ordem 60. Para valores de ordem N superior a 60 os erros das soluções de colocação Δu_N não são inferiores ao erro Δu_{60} e, consequentemente apenas introduzimos ruídos indesejáveis nos coeficientes espectrais.

Observando a localização de zeros e polos dos ACP mistos $\mathcal{H}_m^{(60)}$ verificámos que todos os filtros $\mathcal{H}_m^{(60)}$ disponíveis não possuem pares de Froissart em Ω.

Os erros $\Delta\mathcal{H}_m^{(60)}$ diminuem muito lentamente com o aumento de m e nunca melhoram o erro do filtro espectral Δu_{60}. Na Figura 4.15 comparamos o erro da solução espectral u_{60} com os erros dos filtros $\mathcal{H}_{10}^{(60)}$ e $\mathcal{H}_{19}^{(60)}$. Em termos de erro máximo absoluto, tem-se $\max\limits_{(x,y)\in\Omega} \Delta u_{60} = 3.42e{-}2$, $\max\limits_{(x,y)\in\Omega} \Delta\mathcal{H}_{10}^{(60)} = 5.81e{-}2$ e $\max\limits_{(x,y)\in\Omega} \Delta\mathcal{H}_{19}^{(60)} = 4.01e{-}2$.

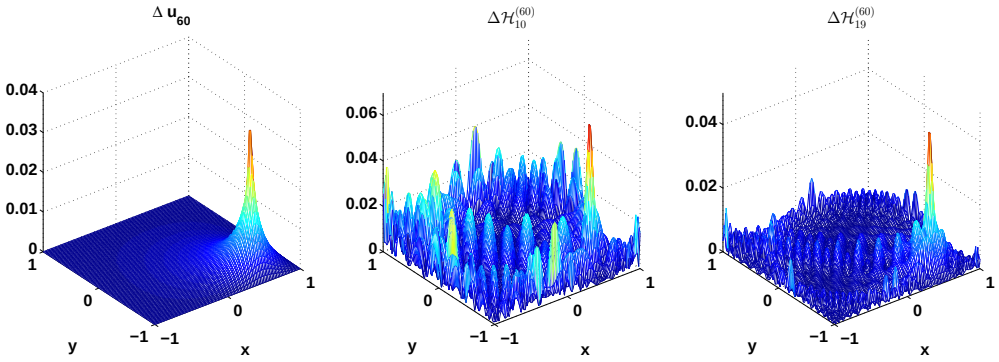


Figura 4.15: Erro absoluto da solução de colocação Δu_{60} (esquerda) Erros absolutos dos filtros mistos $\Delta\mathcal{H}_{10}^{(60)}$ e $\Delta\mathcal{H}_{19}^{(60)}$ (centro e direita) do problema B.

4.5.2 ACP "homogéneos" do tipo I

ACP "homogéneos" do tipo I da função salto

De forma análoga aos exemplos anteriores, testámos o comportamento de ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$ da função salto (4.31). Os resultados obtidos foram idênticos aos obtidos para os AP mistos $\mathcal{H}_m$. Ou seja, todos os AP ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$ observados apresentam pares de Froissart em Ω, o que destrói a qualidade das suas aproximações. Note-se que para estes aproximantes podemos fixar o conjunto de índices do denominador D_m e variar o conjunto de índices do numerador N_n. É interessante verificar que nestas sequências os zeros dos aproximantes tendem a espalhar-se pela região de Ω onde a função salto se anula e os pares de Froissart apenas se localizam na região onde a função toma o valor 1. Ilustramos este facto na figura 4.16 onde apresentamos aproximantes com $m = 1$ fixo e valores de $n = 1, 4, \ldots, 25$. Quando se aumenta o valor de m os pares de Froissart tendem a preencher a região de Ω onde a função salto toma o valor 1.

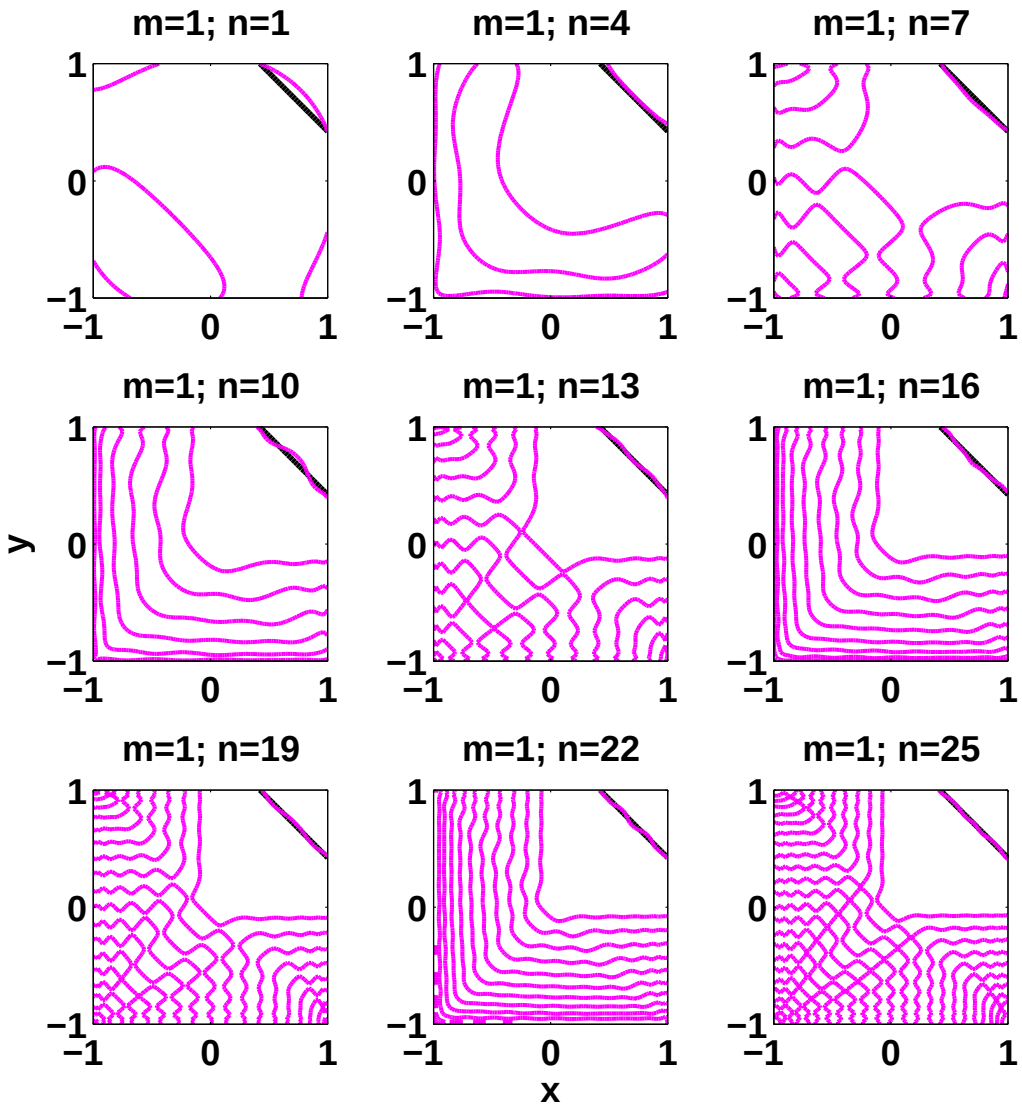


Figura 4.16: Localização de pólos (curva a preto) e de zeros (curvas a magenta) de aproximantes ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$ da função salto com o conjunto de D_1 fixo e com o conjunto N_n a variar.

Filtragem espectral via AP "homogéneos" do tipo I

Os filtros dos problemas A e B via AP "homogéneos" do tipo I têm um comportamento similar à filtragem via AP mistos. Nomeadamente, todos os aproximantes ${}^{\mathrm{I}}\mathfrak{H}^{(60)}_{m,n}$ atingíveis, não possuem pares de Froissart em Ω e os valores máximos de $\Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{m,n}$ em Ω atingem valores próximos dos valores máximos da solução Δu_{60} sem contudo conseguirem melhorar a aproximação espectral. Contudo, tanto no problema A como no

problema B, os melhores filtros ${}^{\mathrm{I}}\mathfrak{H}^{(60)}_{m,n}$ foram obtidos para valores de $m = 1$ e valor de n máximo. Na Figura 4.17 apresentam-se os erros absolutos máximos $\max\limits_{(x,y)\in\Omega} \Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,n}$, para $n = 1, 2, \ldots, 56$ comparados com os erros absolutos máximos $\max\limits_{(x,y)\in\Omega} \Delta u_{60}$ para os problemas A, B, C e D. Como podemos verificar no problema A, o valor máximo do erro da solução espectral é $\max\limits_{(x,y)\in\Omega} \Delta u_{60} = 7.92e{-}2$ e o valor máximo do erro do melhor filtro é $\max\limits_{(x,y)\in\Omega} \Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56} = 7.41e{-}2$. No problema B o valor máximo do erro da solução espectral é $\max\limits_{(x,y)\in\Omega} \Delta u_{60} = 3.62e{-}2$ e o valor máximo do erro do melhor filtro é $\max\limits_{(x,y)\in\Omega} \Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56} = 3.74e{-}2$. Ou seja os resultados dos filtros "homogéneos" do tipo I são melhores, mas não significativamente, do que os resultados obtidos com filtros mistos mas não melhoraram os resultados obtidos pela aproximação espectral.

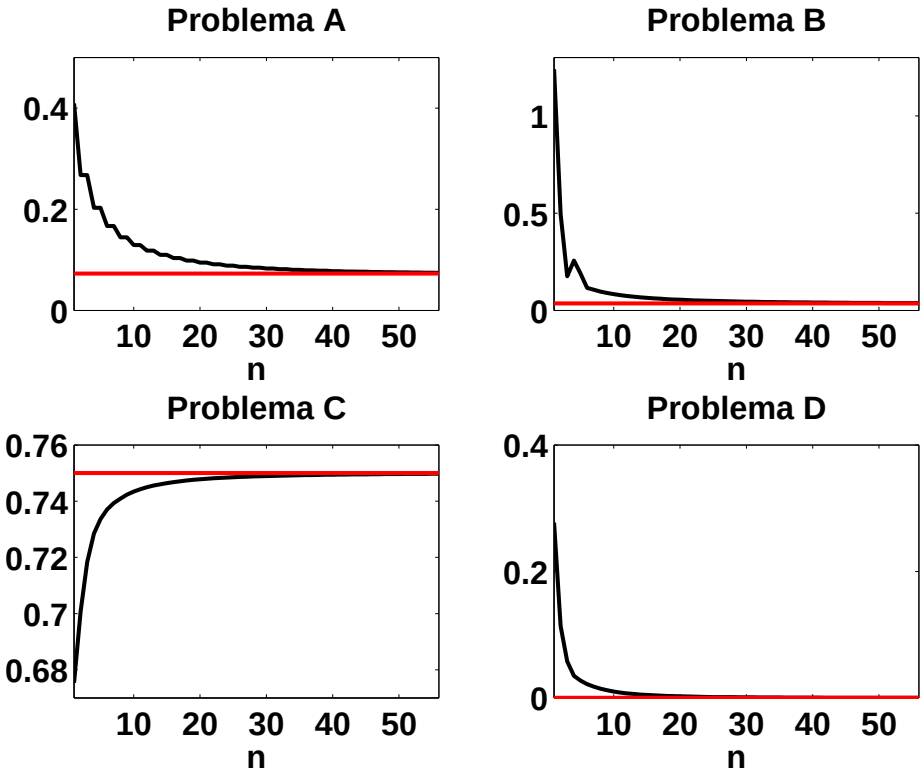


Figura 4.17: Erros absolutos máximos $\max\limits_{(x,y)\in\Omega} {}^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,n}$ para valores de $n = 1, 2, \ldots, 56$ (curvas a preto), erros absolutos máximos $\max\limits_{(x,y)\in\Omega} \Delta u_{60}$ (linha vermelha) para os problemas A, B, C e D.

Iremos, de seguida, considerar a filtragem do problema C, com singularidade $s_C = (1/2, 1)$ na fronteira de Ω e a filtragem do problema D, com singularidade $s_D = (1, 1)$ num vértice da fronteira de Ω. De salientar que para os dois problemas a melhor solução

espectral é a solução de ordem $N = 60$ e que todos os filtros não possuem pólos em Ω.

Filtragem do problema C Neste problema o melhor filtro, no sentido de que minimiza $\max_{(x,y)\in\Omega} \Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{m,n}$ é o filtro de ordem mais baixa ${}^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,1}$, ver Figura 4.17. Na realidade tem-se $\max_{(x,y)\in\Omega} {}^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,1}(x,y) = 6.75e-1$ que é ligeiramente inferior ao valor máximo do erro absoluto da solução espectral $\max_{(x,y)\in\Omega} u_{60}(x,y) = 7.5e-1$. No entanto, sacrificamos a aproximação dada por ${}^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,1}$ nos pontos de Ω mais afastados da singularidade, ver Figura 4.18.

Uma última observação, dado que a singularidade s_C se encontra no lado do quadrado $y = 1$ da fronteira de Ω e que a função que define as condições fronteira é a função h_W, tentou-se melhorar os resultados aumentando a ordem da solução de primeira ordem para calcular os coeficientes de h_W com maior precisão, ver exemplo 1.4.2. Contudo os resultados obtidos não sofreram quaisquer alterações.

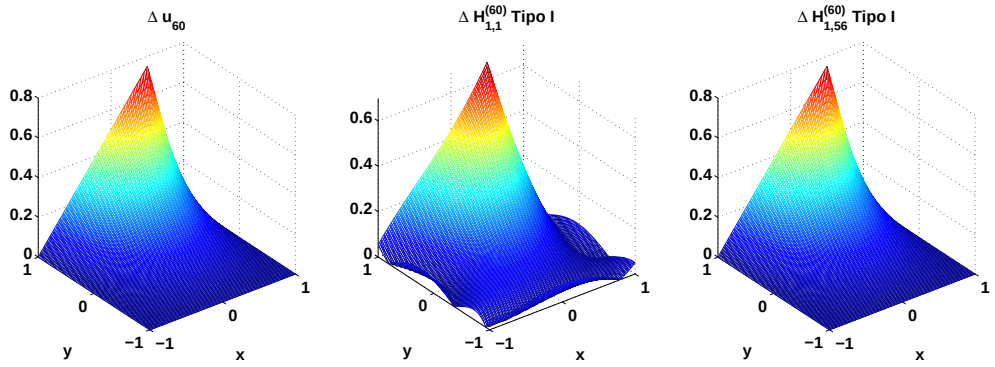


Figura 4.18: Erros absolutos do problema C. (esquerda) Δu_{60}, (centro) $\Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,1}$, (direita) $\Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56}$.

Filtragem do problema D Entre os quatro problemas tratados, o problema D é o que exibe convergência do método de colocação mais rápida. De facto, observou-se que $\max_{(x,y)\in\Omega} \Delta u_{60} = 3.99e-6$. O filtro com melhor aproximação foi, analogamente aos problemas A e B, o aproximante ${}^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56}$, com $\max_{(x,y)\in\Omega} {}^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56} = 1.60e-4$. Na Figura 4.19 representamos, em escala logarítmica, os erros Δu_{60}, $\Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56}$ e $\Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{19,19}$. Note-se que o

erro absoluto do aproximante ${}^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56}$ é inferior ao erro absoluto do aproximante ${}^{\mathrm{I}}\mathfrak{H}^{(60)}_{19,19}$ em quase todos os pontos de Ω.

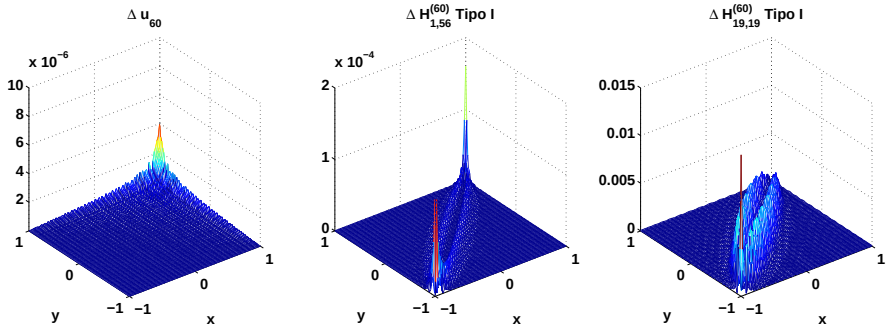


Figura 4.19: Erros absolutos do problema D. (esquerda) Δu_{60}, (centro) $\Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{1,56}$, (direita) $\Delta^{\mathrm{I}}\mathfrak{H}^{(60)}_{19,19}$.

4.5.3 ACP "homogéneos" do tipo II

Começamos por observar que enquanto nos ACP "homogéneos" do tipo I existem sequências de aproximantes com o conjunto D_m fixo, $\left\{{}^{\mathrm{I}}\mathfrak{H}_{m,n}\right\}_{n\geq 1}$, com m fixo, nos ACP "homogéneos" do tipo II existem sequências de aproximantes com o conjunto N_m fixo, $\left\{{}^{\mathrm{II}}\mathfrak{H}_{m,n}\right\}_{m\geq 1}$, com n fixo.

ACP "homogéneos" do tipo II da função salto

Relativamente à localização dos pólos, verificou-se que todos os AP "homogéneos" do tipo II observados, possuem pólos espúrios e/ou pares de Froissart em Ω, com a exceção de dois casos, ${}^{\mathrm{II}}\mathfrak{H}_{6,1}$ e ${}^{\mathrm{II}}\mathfrak{H}_{8,1}$, ver Figura 4.20. Os erros para estes dois últimos, encontram-se ilustrados na figura 4.21 na qual se representam os erros da série truncada (4.33) $\Delta f_{8,1}$ e os erros ${}^{\mathrm{II}}\mathfrak{H}_{6,1}$ e ${}^{\mathrm{II}}\mathfrak{H}_{8,1}$. Estas aproximações de Padé não melhoram a aproximação dada pela série truncada $f_{8,1}$.

Filtragem espectral via AP "homogéneos" do tipo II

A filtragem dos quatro problemas com filtros "homogéneos" do tipo II revelou resultados em tudo semelhantes aos resultados descritos acima, para filtros "homogéneos" do

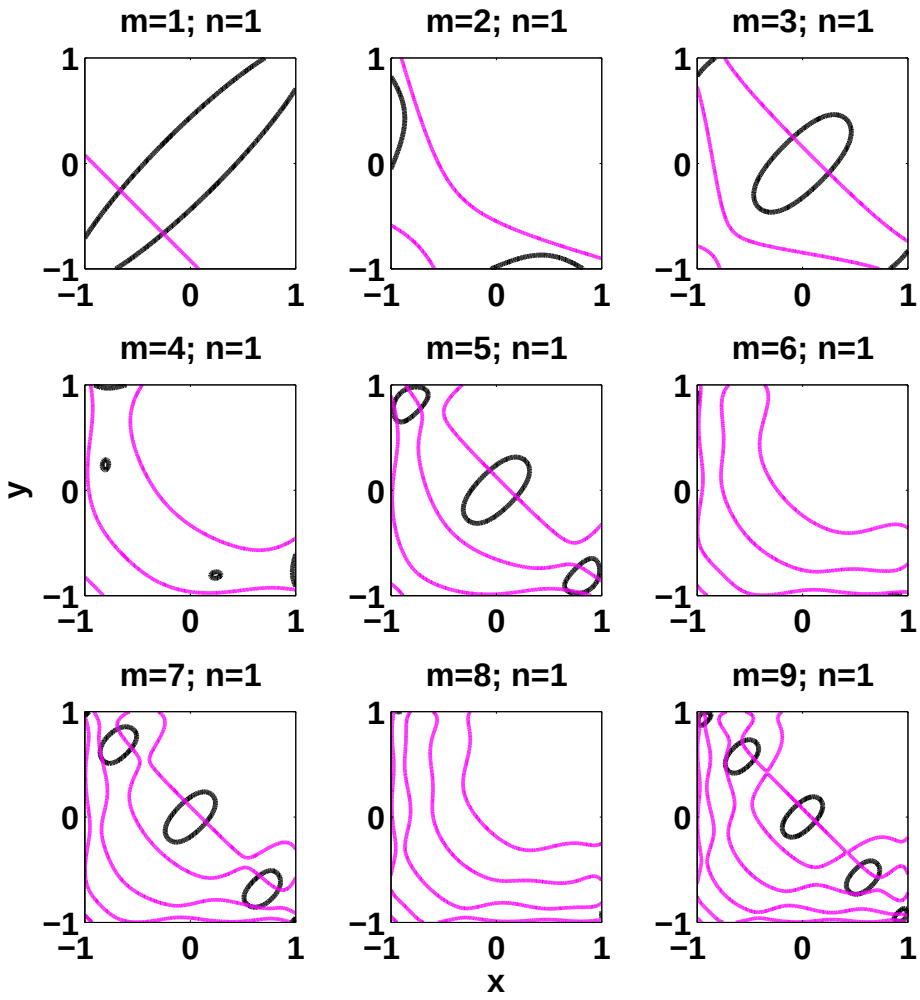


Figura 4.20: Localização dos pólos (curvas a preto) e dos zeros (curvas a magenta) dos ACP ${}^{\text{II}}\mathfrak{H}_{m,1}$, $m = 1, 2, \ldots, 9$, da função salto.

tipo I. A principal diferença deve-se ao facto de que para uma solução espectral de ordem N fixa, o número de AP do tipo II acessíveis não é necessariamente igual ao número de AP do tipo I acessíveis. Por exemplo para uma solução espectral de ordem $N = 60$, tem-se para os aproximantes ${}^{\text{I}}\mathfrak{H}_{1,n}^{(60)}$ estão acessíveis para valores de $n = 1, 2, \ldots, 56$, enquanto os AP do tipo II ${}^{\text{II}}\mathfrak{H}_{1,n}^{(60)}$ apenas estão disponíveis para valores de $n = 1, 2, \ldots, 28$.

Em jeito de resumo apresentamos na tabela 4.1 os valores máximos dos erros absolutos da solução de colocação u_{60} e dos três aproximantes de equações provenientes de reticulados,$\mathcal{H}_m$, ${}^{\text{I}}\mathfrak{H}_{m,n}$ e ${}^{\text{II}}\mathfrak{H}_{m,n}$, para os quatro problemas estudados. Podemos observar que, com a exceção do problema C, os aproximantes homogéneos do tipo I forneceram a melhor aproximação de Padé mas não melhoram a aproximação de colocação. Para o

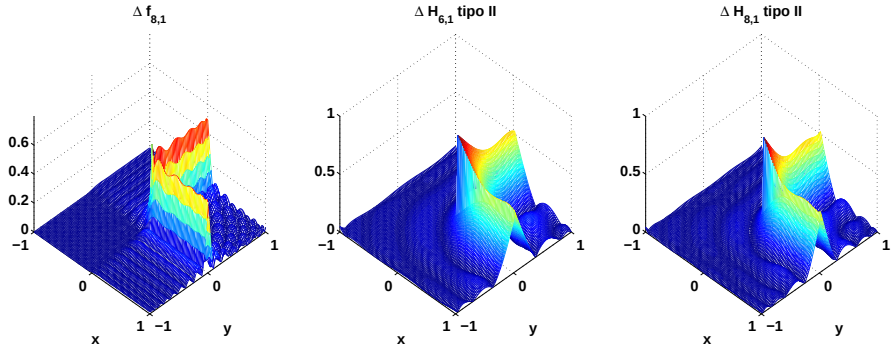


Figura 4.21: Erros absolutos: $\Delta f_{8,1}$ (esquerda), $\Delta^{\mathrm{II}}\mathfrak{H}_{6,1}$ (centro) e $\Delta^{\mathrm{II}}\mathfrak{H}_{8,1}$ (direita) da função salto.

problema C todos os filtros melhoram a aproximação dada pela solução de colocação u_{60}, e o filtro que fornece a melhor aproximação é o filtro homogéneo do tipo II. Todavia, esta melhoria é apenas local, numa região próxima da singularidade s_C que se situa no ponto médio de um dos lados que delimita o domínio Ω.

	$\max\limits_{(x,y)\in\Omega} \Delta u_{60}$	$\max\limits_{(x,y)\in\Omega} \Delta\mathcal{H}_m$	$\max\limits_{(x,y)\in\Omega} \Delta^{\mathrm{I}}\mathfrak{H}_{m,n}$	$\max\limits_{(x,y)\in\Omega} \Delta^{\mathrm{II}}\mathfrak{H}_{m,n}$
Problema A	$7.92e{-2}$	$8.12e{-2}$ $m = 19$	$7.41e{-2}$ $m = 1,\ n = 56$	$8.63e{-2}$ $m = 1,\ n = 28$
Problema B	$3.62e{-2}$	$4.01e{-2}$ $m = 19$	$3.74e{-2}$ $m = 1,\ n = 56$	$4.74e{-2}$ $m = 1,\ n = 28$
Problema C	$7.50e{-1}$	$6.54e{-1}$ $m = 1$	$6.75e{-1}$ $m = 1,\ n = 1$	$6.35e{-1}$ $m = 1,\ n = 1$
Problema D	$3.99e{-6}$	$3.12e{-2}$ $m = 19$	$1.60e{-4}$ $m = 1,\ n = 56$	$6.53e{-4}$ $m = 1,\ n = 28$

Tabela 4.1: Erros absolutos máximos das soluções espectrais u_{60} e dos melhores filtros $\mathcal{H}_m$, ${}^{\mathrm{I}}\mathfrak{H}_{m,n}$ e ${}^{\mathrm{II}}\mathfrak{H}_{m,n}$ para os quatro problemas considerados. Os valores de m e de n indicados representam a ordem do melhores filtros.

4.6 Observações e conclusões

Numa primeira observação notamos que não encontrámos na literatura resultados sobre convergência de aproximantes de Padé de séries ortogonais multidimensionais. Deste modo todas as conclusões efetuadas nesta secção baseiam-se apenas nos resultados obtidos nos diversos testes numéricos efetuados. A segunda observação reside no facto dos resultados obtidos nas duas abordagens de ACP bidimensionais (*nested* e provenientes de reticulados) dependerem fortemente das funções a aproximar. Além disso os resultados obtidos com ACP provenientes de reticulados dependem também da geometria dos conjuntos de índices N, D e E. Deste modo, dada uma função bidimensional f, o problema da escolha de um "bom" ACP bidimensional de f não é um problema trivial.

Para explicar os resultados obtidos iremos considerar duas classes de funções em $L_\omega([-1,1]^2)$: $\mathcal{F}_1$ a classe das funções que possuem um conjunto de singularidades não isoladas em Ω, e $\mathcal{F}_2$ a segunda classe de funções constituída por funções analíticas ou por funções que possuem um conjunto de singularidades isoladas em Ω. Como o nosso objetivo é usar os AP como filtros de soluções espectrais de problemas diferenciais rígidos (stiff) estudamos sobretudo o comportamento de AP de funções da classe $\mathcal{F}_1$ e de funções da classe $\mathcal{F}_2$ com singularidades isoladas no domínio Ω.

ACP de funções em $\mathcal{F}_1$

No caso das funções em $\mathcal{F}_1$, os testes numéricos efetuados revelaram que os aproximantes *nested* possuem algumas vantagens sobre os aproximantes de equações provenientes de reticulados. De facto, se analisarmos os resultados obtidos para a função salto, observamos que todos os ACP provenientes de reticulados possuem pólos espúrios e/ou pares de Froissart em Ω, com a exceção de ${}^{\text{II}}\mathfrak{H}_{6,1}$ e de ${}^{\text{II}}\mathfrak{H}_{8,1}$. Ao invés, os aproximantes *nested* $R_{m,n}$, da função salto não exibem pólos espúrios e/ou pares de Froissart em Ω. Consequentemente parece ser preferível optar pelos ACP *nested*, ou pelos ACP homogéneos do tipo II, ao aproximar funções da classe $\mathcal{F}_1$. Todavia notamos que, geralmente, o uso de ACP bidimensionais, de funções em $\mathcal{F}_1$, não melhora os resultados dados pelas soluções espectrais.

ACP de funções em $\mathcal{F}_2$

Para funções em $\mathcal{F}_2$ com singularidades isoladas no domínio Ω, justifica-se o uso dos filtros definidos por equações provenientes de reticulados apresentados, sendo que, se obteve melhores resultados com filtros homogéneos do tipo II. Todavia estes filtros apenas melhoram localmente, e em certos casos especiais, a aproximação dada pela solução espectral. Note-se que nos exemplos apresentados nas secções anteriores apenas no problema C, com singularidade no ponto médio de um dos lados da fronteira de Ω, os filtros melhoram

a aproximação da solução espectral, mas apenas em pontos relativamente próximos da singularidade s_C.

Analogamente ao caso unidimensional, ver Exemplo 3.3.2, se a solução de uma equação diferencial tiver "mudanças bruscas de valores" então a convergência exibida pelos métodos espectrais é lenta. Note-se que estas funções tanto podem pertencer à classe $\mathcal{F}_1$ ou à classe $\mathcal{F}_2$. Iremos de seguida apresentar os resultados obtidos do seguinte exemplo que é uma versão bidimensional do Exemplo 3.3.2.

Exemplo 4.6.1. Consideramos, neste exemplo, novamente a equação de Poisson bidimensional, com as funções g, h_W, h_E, h_S e h_N definidas respetivamente por

$$g(x,y) = -\frac{8(x+y)}{\epsilon^3\sqrt{\pi}} e^{-\frac{(x+y)^2}{\epsilon^2}} \tag{4.41}$$

$$h_W(y) = \text{erf}\left(\frac{y-1}{\epsilon}\right) \tag{4.42}$$

$$h_E(y) = \text{erf}\left(\frac{y+1}{\epsilon}\right) \tag{4.43}$$

$$h_N(x) = \text{erf}\left(\frac{x+1}{\epsilon}\right) \tag{4.44}$$

$$h_S(x) = \text{erf}\left(\frac{x-1}{\epsilon}\right) \tag{4.45}$$

onde, $\epsilon \in \mathbb{R}^+$ e erf é a função erro.

A solução deste problema é a função u definida por

$$u(x,y) = \text{erf}\left(\frac{x+y}{\epsilon}\right)$$

que se anula na reta $x+y=0$. Quando ϵ tende para zero a função u tende pontualmente para a função descontínua s definida por

$$s(x,y) = \begin{cases} 1, & x+y>0 \\ 0, & x+y=0 \\ -1, & x+y<0 \end{cases}.$$

Deste modo, a solução u, para valores de ϵ próximos de zero, sofre uma mudança brusca numa vizinhança da reta $x+y=0$. Resolvendo este problema, com $\epsilon = 10^{-2}$, usando o método de colocação obteve-se a solução de colocação u_N, $N=75$, com erro máximo absoluto em Ω, $\max_{(x,y)\in\Omega} \Delta u_{75}(x,y) = 9.79e-1$. Para valores de $N > 75$ as soluções de colocação não melhoram significativamente este resultado. Os resultados obtidos são ilustrados na Figura 4.22.

Foram testados os quatro filtros considerados neste trabalho, e observou-se que os filtros *nested*, mistos e homogéneos do tipo I possuíam todos pólos espúrios ou pares de

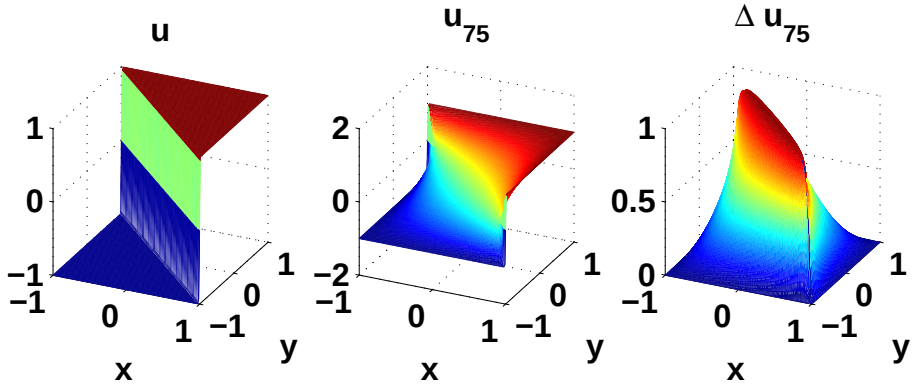


Figura 4.22: Solução do exemplo 4.6.1 com $\epsilon = 10^{-2}$ (esquerda), Solução espectral u_{75} (centro), Erro absoluto da solução espectral Δu_{75} (direita).

Froissart no domínio Ω. Todavia existem sucessões de filtros homogéneos do tipo I, ${}^{\mathrm{II}}\mathfrak{H}^{(75)}_{m,n}$ isentos de pólos espúrios/pares de Froissart no domínio Ω. Mais exatamente, se fixarmos o valor de m_0, e escolhermos um valor $n_0 > m_0$ suficientemente grande então todos os filtros ${}^{\mathrm{II}}\mathfrak{H}^{(75)}_{m_0,n}$, $n \geq n_0$ acessíveis não possuem pólos espúrios nem pares de Froissart no domínio Ω. Este facto mostra que além da escolha dos ACP bidimensionais ser crucial, também a escolha dos conjuntos N_n, D_m e E_k é igualmente importante no processo de filtragem de problemas bidimensionais. Se fixarmos o valor $m_0 = 2$, observamos que para valores de n ímpares tais que $n \geq n_0$, $n_0 = 9$, os filtros ${}^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,n}$ acessíveis, $9 \leq n \leq 35$ não possuem pólos espúrios nem pares de Froissart no domínio Ω. Apresentamos a localização dos pólos e zeros de alguns filtros da sequência ${}^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,n}$, $n \geq 1$ na Figura 4.23.

Na Figura 4.24 representamos, com uma linha vermelha, o erro absoluto máximo da solução de colocação e, com quadrados, os erros absolutos máximos dos filtros

$$\max_{(x,y)\in\Omega} \Delta^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,n}(x,y), \quad n = 9, 11, \ldots, 35.$$

Podemos observar que apenas o filtro ${}^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,33}$ melhora ligeiramente o erro máximo da solução de colocação. Na realidade, os gráficos indicados na Figura 4.25 mostram que o

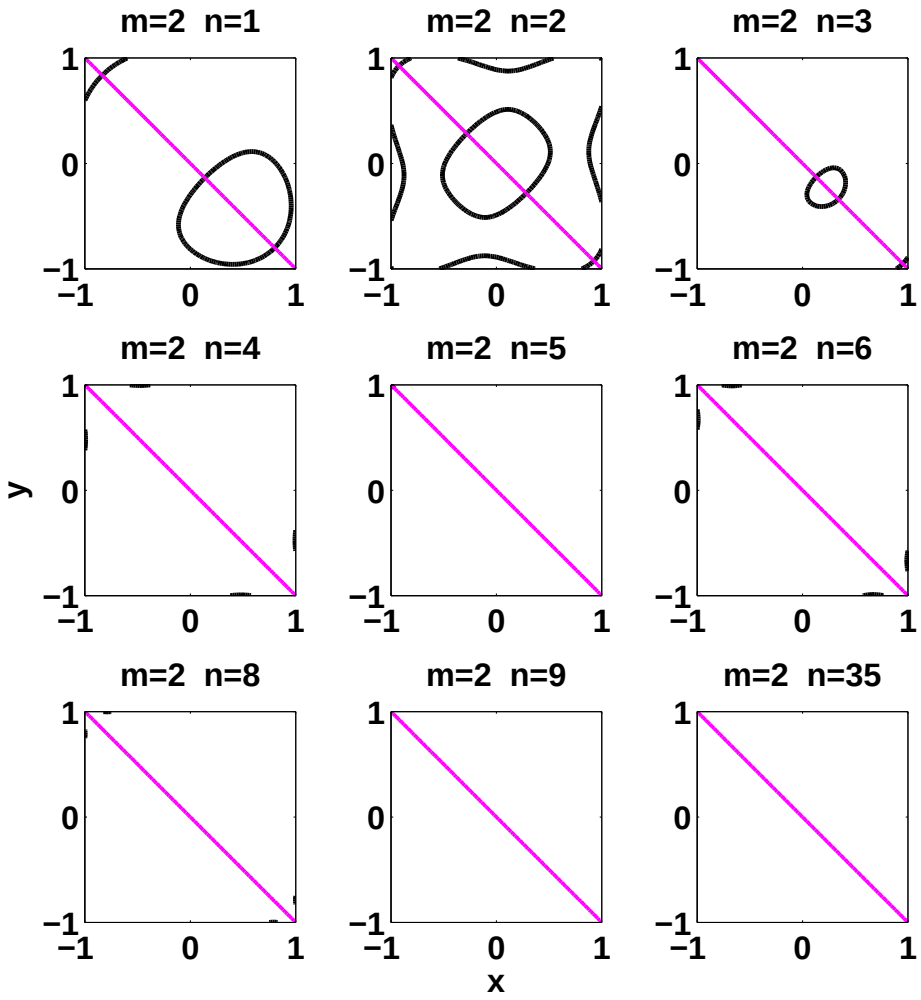


Figura 4.23: Localização dos zeros (curvas de cor magenta) e pólos (curvas de cor preta) de alguns filtros da sequência ${}^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,n}$, do Exemplo 4.6.1 com $\epsilon = 10^{-2}$.

gráfico do erro Δu_{75} é quase indistinguível do gráfico do erro $\Delta^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,33}$, com a exceção de pontos situados junto da fronteira do domínio Ω. Note-se que nos pontos da fronteira de Ω a aproximação dada pelo método de colocação é naturalmente melhor que as aproximações de Padé, dado que as soluções de colocação coincidem com a solução nos nós situados na fronteira de Ω enquanto os AP não.

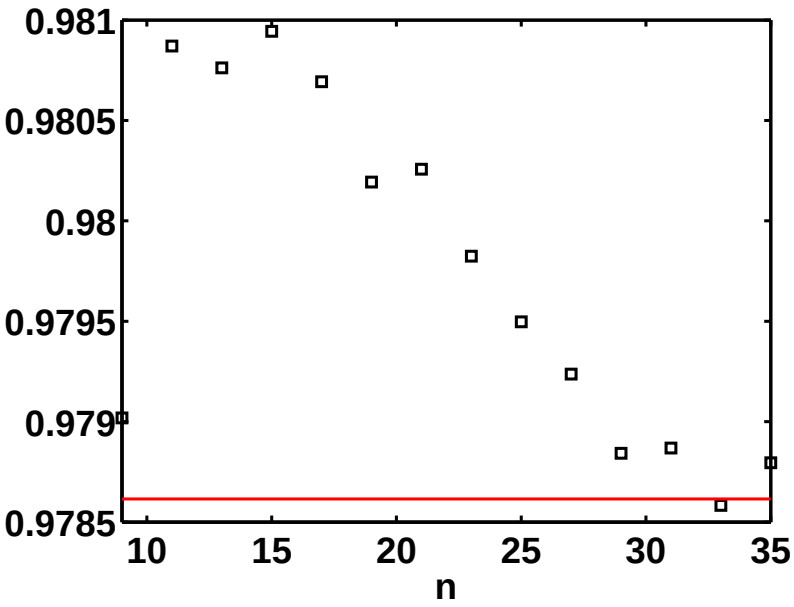


Figura 4.24: $\max_{(x,y)\in\Omega} \Delta^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,n}(x,y)$, $n = 9, 11, \ldots, 35$, (assinalados com um quadrado). Erro absoluto máximo da solução de colocação $\max_{(x,y)\in\Omega} \Delta u_{75}(x,y) = 9.78795e-1$ do Exemplo 4.6.1 com $\epsilon = 10^{-2}$ (linha vermelha).

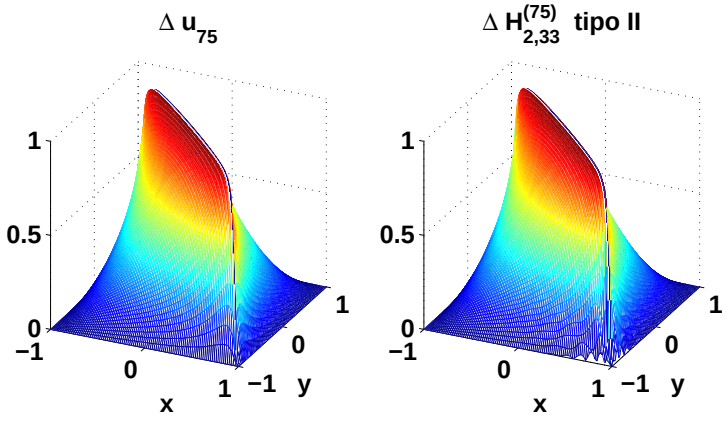


Figura 4.25: Erro absoluto da solução de colocação u_{75} (Esquerda), erro absoluto do filtro $\Delta^{\mathrm{II}}\mathfrak{H}^{(75)}_{2,33}$ (Direita).

Apêndice A

Polinómios ortogonais

A.1 Problemas de Sturm-Liouville

A importância dos problemas de Sturm-Liouville (S-L) nos métodos espectrais reside no facto das aproximações espectrais serem combinações lineares de funções próprias, também conhecidas por funções base, de um problema de Sturm-Liouville. Além disso a taxa de convergência de um método espectral está intimamente relacionada com o espectro do problema de S-L associado às funções base escolhidas. De facto, o nome dado a estes métodos, *Métodos espectrais*, deve-se a esta característica. Abordaremos com mais detalhe estas características, dos métodos espectrais, no apêndice C.

Um problema de Sturm-Liouville é um problema da forma [CH89]

$$-(pu')' + qu = \lambda wu, \quad \text{em } I, \quad \lambda \in \mathbb{R} \tag{A.1}$$

onde I é um intervalo aberto em $\mathbb{R}$ e as funções $p : \bar{I} \longrightarrow \mathbb{R}$, $q : I \longrightarrow \mathbb{R}$ e $w : I \longrightarrow \mathbb{R}$ são contínuas com $p \geq 0$ em $\bar{I}$ e $w > 0$ em I. Se $p > 0$ em $\bar{I}$ o problema diz-se *regular*, e, no caso de p ter pelo menos um zero o problema diz-se *singular*.

Nos métodos espectrais estamos interessados nos problemas para os quais as soluções $\{(\lambda_n, u_n)\}_{n\in\mathbb{N}_0}$ verificam $\lambda_n > 0$, $\lim\limits_{n\to\infty} \lambda_n = \infty$, onde u_n são polinómios de grau n devidamente normalizados. As funções próprias polinomiais de um problema (A.1) satisfazem uma relação de recorrência a dois termos. Mais exactamente temos o seguinte

Teorema A.1.1. Se $\{u_n\}_{n\in\mathbb{N}_0}$ é uma sucessão de soluções de (A.1), onde u_n é um polinómio de grau n, então é possível encontrar três sucessões de números reais: $\{\rho_n\}_{n\geq 2}$, $\{\sigma_n\}_{n\geq 2}$ e $\{\xi_n\}_{n\geq 2}$ tais que,

$$u_n(x) = (\rho_n x + \sigma_n)u_{n-1}(x) + \xi_n u_{n-2}(x) \quad \forall n \geq 2, \ \forall x \in I. \tag{A.2}$$

Definindo ρ_1 e σ_1 de forma a que $u_1(x) = (\rho_1 x + \sigma_1)u_0(x)$ podemos calcular uma solução polinomial de grau n, usando a relação (A.2) a partir da solução polinomial de grau zero.

Derivando (A.2) encontramos uma relação para as derivadas das soluções polinomiais

$$u_n'(x) = (\rho_n x + \sigma_n)u_{n-1}'(x) + \rho_n u_{n-1}(x) + \xi_n u_{n-2}'(x) \quad \forall n \geq 2,\ \forall x \in I, \tag{A.3}$$

com condições iniciais $u_0'(x) = 0$ e $u_1'(x) = \rho_1 u_0$. De modo análogo podemos encontrar relações de recorrência para as derivadas de ordem superior.

Iremos ver de seguida as famílias de soluções do problema (A.1) mais relevantes para os métodos espectrais.

A.2 Polinómios de Jacobi

Considerando em (A.1) $I =]-1, 1[$, $p(x) = (1-x)^{\alpha+1}(1+x)^{\beta+1}$, $\forall x \in \bar{I}$, $q(x) = 0$ e $w(x) = (1-x)^{\alpha}(1+x)^{\beta}$, $\forall x \in I$, onde $\alpha, \beta > -1$ obtemos o seguinte problema de Sturm-Liouville

$$-(1-x^2)u'' + ((\alpha+\beta+2)x + \alpha - \beta)u' = \lambda u. \tag{A.4}$$

Aplicando o método de Frobenius à equação (A.4), obtemos o seguinte resultado,

Teorema A.2.1. Se uma solução de (A.4) é um polinómio de grau n então, o seu valor próprio associado λ_n é dado por $\lambda_n = n(n+\alpha+\beta+1)$, $n = 0, 1, 2 \ldots$.

A cada valor próprio da forma $\lambda_n = n(n+\alpha+\beta+1)$, $n \in \mathbb{N}_0$, estão associados funções próprias polinomiais de grau n que se distinguem apenas por um fator multiplicativo. Definimos os polinómios de Jacobi, $P_n^{(\alpha,\beta)}$ como sendo a única função própria polinomial de grau n normalizado pela condição,

$$P_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n} = \frac{\Gamma(n+\alpha+1)}{n!\Gamma(\alpha+1)}, \quad n \in \mathbb{N}_0, \tag{A.5}$$

ou, equivalentemente, por

$$P_n^{(\alpha,\beta)}(-1) = (-1)^n \binom{n+\beta}{n} \tag{A.6}$$

onde Γ é a função gama, definida, para todo o x real positivo, por

$$\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}dt.$$

Existem muitos resultados para a família dos polinómios de Jacobi, ver p. ex. [Ask75]. Resumimos de seguida alguns resultados sobre os polinómios de Jacobi úteis para os métodos espectrais. (fórmula de Rodrigues)

$$(1-x)^{\alpha}(1+x)^{\beta}P_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}\frac{d^n}{dx^n}\left\{(1-x)^{n+\alpha}(1+x)^{n+\beta}\right\} \tag{A.7}$$

Os polinómios de Jacobi verificam igualmente a seguinte fórmula

$$\begin{aligned} P_n^{(\alpha,\beta)}(x) &= \frac{1}{2^n}\sum_{k=0}^{n}\binom{n+\alpha}{k}\binom{n+\beta}{n-k}(x-1)^{n-k}(x+1)^k \\ &= \frac{\Gamma(2n+\alpha+\beta+1)}{2^n n!\Gamma(n+\alpha+\beta+1)}\left[x^n + \frac{(\alpha-\beta)n}{2n+\alpha+\beta}x^{n-1}+\cdots\right], \quad n\in\mathbb{N}_0. \end{aligned} \tag{A.8}$$

Para os polinómios de Jacobi a relação (A.2) toma a forma

$$P_n^{(\alpha,\beta)} = (\rho_n x + \sigma_n)P_{n-1}^{(\alpha,\beta)} + \xi_n P_{n-1},\ n\geq 2 \tag{A.9}$$

onde

$$P_0^{(\alpha,\beta)}(x) = 1, \quad P_1^{(\alpha,\beta)}(x) = \frac{1}{2}(\alpha+\beta+2)\,x + \frac{1}{2}(\alpha-\beta), \tag{A.10}$$

sendo as sucessões $\{\rho_n\}_{n\geq 2}$, $\{\sigma_n\}_{n\geq 2}$ e $\{\xi_n\}_{n\geq 2}$ determinadas por

$$\begin{aligned} \rho_n &= \frac{(2n+\alpha+\beta)(2n+\alpha+\beta-1)}{2n(n+\alpha+\beta)} \\ \sigma_n &= \frac{(\alpha^2-\beta^2)(2n+\alpha+\beta-1)}{2n(n+\alpha+\beta)(n+\alpha+\beta-2)} \\ \xi_n &= -\frac{(n+\alpha-1)(n+\beta-1)(2n+\alpha+\beta)}{n(n+\alpha+\beta)(2n+\alpha+\beta-2)}, \quad n\geq 2 \end{aligned} \tag{A.11}$$

E, no intervalo $\bar{I}$, tem-se [Sze39]

$$\max_{x\in[-1,1]}\left|P_n^{(\alpha,\beta)}(x)\right| = \max\left\{\left|P_n^{(\alpha,\beta)}(\pm 1)\right|\right\} = \max\left\{\binom{n+\alpha}{n}, \binom{n+\beta}{n}\right\}, \quad n\in\mathbb{N}. \tag{A.12}$$

Diferentes escolhas dos parâmetros α e β originam diferentes famílias de polinómios de Jacobi. Se escolhermos valores de α e β tais que $\alpha=\beta$ obtemos a família dos chamados polinómios *ultra esféricos*, também chamados de polinómios de Gegenbauer. Na família dos polinómios de Gegenbauer, destacamos os polinómios de Legendre e os polinómios de Chebyshev.

A.2.1 Polinómios de Legendre

Os polinómios de Legendre são polinómios de Jacobi ultra esféricos com $\alpha=\beta=0$. São representados por P_n (em vez de $P_n^{(0,0)}$) e são funções próprias do problema de Sturm-Liouville

$$-\left((x^2-1)u'\right)' = n(n+1)u \tag{A.13}$$

As condições (A.5) e (A.6) para os polinómios de Legendre são

$$P_n(1) = 1, \qquad P_n(-1) = (-1)^n, \quad n\in\mathbb{N}. \tag{A.14}$$

E a sua relação de recorrência toma a forma

$$P_n(x) = \frac{2n-1}{n}xP_{n-1}(x) - \frac{n-1}{n}P_{n-2}(x), \quad x\in\bar{I}, \quad n\geq 2. \tag{A.15}$$

com $P_0(x) = 1$ e $P_1(x) = x$. Uma importante propriedade é que P_n é uma função par se n for par e é uma função ímpar se n for ímpar. Outras propriedades de interesse para os métodos espectrais (para mais detalhes ver [Sze39]) são: de (A.13) pode-se concluir que

$$P_n'(\pm 1) = \pm \frac{n(n+1)}{2}(\pm 1)^n, \quad n \in \mathbb{N}, \tag{A.16}$$

e de (A.12) concluímos que

$$|P_n(x)| \leq 1, \quad x \in \bar{I}, n \in \mathbb{N}. \tag{A.17}$$

Para $n \geq 2$ é ainda possível mostrar que os polinómios de Legendre são uniformemente limitados, em $\bar{I}$, pelas parábolas $y = -\frac{1}{2}(1+x^2)$ e $y = \frac{1}{2}(1+x^2)$. Usando a relação (A.15) podemos concluir que

$$\lim_{n\to\infty} P_n(0) = 0$$

dado que se tem

$$P_n(0) = \begin{cases} 0 & \text{se } n \text{ par}, \\ n! 2^{-n} \left[\left(\frac{n}{2}\right)!\right]^{-2} & \text{se } n \text{ impar}. \end{cases} \tag{A.18}$$

A.2.2 Polinómios de Chebyshev

Os polinómios de Chebyshev de *primeira espécie*, T_n, estão relacionados com os polinómios ultra esféricos pela escolha dos parâmetros $\alpha = \beta = -\frac{1}{2}$ e são definidos por (para mais detalhes ver p. ex. [Luk69])

$$T_n = k_n P_n^{(-\frac{1}{2}, -\frac{1}{2})}, \quad n \in \mathbb{N}_0, \tag{A.19}$$

onde

$$k_n = \frac{(n!2^n)^2}{(2n)!} = \frac{n!\sqrt{\pi}}{\Gamma\left(n+\frac{1}{2}\right)} = \left[\binom{n-\frac{1}{2}}{n}\right]^{-1}.$$

Os polinómios de Chebyshev são funções próprias do problema de Sturm-Liouville

$$\left(\sqrt{1-x^2}u'\right)' + \frac{n^2}{\sqrt{1-x^2}}u = 0, \tag{A.20}$$

e satisfazem a relação de recorrência

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), \quad x \in \bar{I}, \quad n \geq 2, \tag{A.21}$$

onde $T_0 = 1$ e $T_1 = x$.

Uma expressão explicita para os polinómios de Chebyshev de grau n é dada por

$$\begin{aligned} T_n(x) &= \sum_{k=0}^{\left[\frac{n}{2}\right]} \left((-1)^k \sum_{m=k}^{\left[\frac{n}{2}\right]} \binom{n}{2m}\binom{m}{k} \right) x^{n-2k} \\ &= 2^{n-1}x^n - n2^{n-3}x^{n-2} + \tfrac{1}{2}n(n-3)2^{n-5}x^{n-4} + \cdots, \quad x \in \bar{I}, \quad n \in \mathbb{N}_0, \end{aligned} \tag{A.22}$$

consequentemente T_n é uma função par (impar) se n fôr par (impar).

Efetuando $x = \cos\theta$ obtém-se a relação

$$T_n(\cos\theta) = \cos n\theta, \quad \theta \in [0, \pi], \quad n \in \mathbb{N}_0, \tag{A.23}$$

logo, são válidas as seguintes propriedades:

$$T_n(\pm 1) = (\pm 1)^n, \tag{A.24}$$

$$T_n(\pm 1) = (\pm 1)^{n+1} n^2, \tag{A.25}$$

$$|T_n(x)| \leq 1, \quad x \in \bar{I}, \quad n \in \mathbb{N}_0, \tag{A.26}$$

$$|T_n'(x)| \leq n^2, \quad x \in \bar{I}, \quad n \in \mathbb{N}_0, \tag{A.27}$$

$$T_n = \frac{T_{n+1}'}{2(n+1)} - \frac{T_{n-1}'}{2(n-1)}, \quad n \geq 2 \tag{A.28}$$

Os polinómios de Chebyshev de *segunda espécie*, U_n, definem-se como sendo

$$U_n = \frac{1}{n+1} T_{n+1}', \quad n \in \mathbb{N}_0. \tag{A.29}$$

Usando (A.4) e (A.19) temos que os polinómios de Chebyshev de segunda espécie satisfazem

$$U_n = k_n P_n^{(\frac{1}{2}, \frac{1}{2})} \tag{A.30}$$

e, deste modo, pertencem igualmente à família dos polinómios ultra-esféricos. Os polinómios de Chebyshev de segunda espécie satisfazem propriedades semelhantes às propriedades vistas para os de primeira espécie.

Os polinómios de Legendre e os polinómios de Chebyshev (de primeira e segunda ordem) são apropriados para serem usados em aproximações de funções em intervalos limitados. Iremos ver em seguida famílias polinomiais para aproximar funções em intervalos não limitados.

A.3 Polinómios de Laguerre

Seja $\alpha > -1$ e $I =]\,0, +\infty\,[$. Definindo em (A.1) $p(x) = x^{\alpha+1}e^{-x}$, $\forall x \in \bar{I}$, $q(x) = 0$ e $w(x) = x^\alpha e^{-x}$, $\forall x \in I$, temos um problema de Sturm-Liouville da forma

$$xu'' + (\alpha + 1 - x)u' + \lambda u = 0. \tag{A.31}$$

Este problemas apenas admite soluções polinomiais se $\lambda = n$, $n \in \mathbb{N}$.

Definimos o polinómio de Laguerre, L_n^α, de grau n como sendo a (única) função própria do problema (A.31) que satisfaz a condição de normalização

$$L_n^\alpha(0) = \binom{n+\alpha}{n}, \quad n \in \mathbb{N}_0, \quad \alpha > -1. \tag{A.32}$$

Para os polinómios de Laguerre a fórmula de Rodrigues fica

$$e^{-x}x^{\alpha}L_n^{(\alpha)}(x) = \frac{1}{n!}\frac{d^n}{dx^n}\left(e^{-x}x^{n+\alpha}\right), \tag{A.33}$$

é válida a igualdade

$$L_n^{(\alpha)}(x) = \sum_{k=0}^{n}\binom{n+\alpha}{n-k}\frac{(-1)^k}{k!}x^k, \tag{A.34}$$

e a relação de recorrência (A.2) toma a forma

$$L_n^{(\alpha)}(x) = \frac{2n+\alpha-1-x}{n}L_{n-1}^{(\alpha)}(x) - \frac{n+\alpha-1}{n}L_{n-2}^{(\alpha)}(x), \quad \forall n \geq 2, \tag{A.35}$$

onde $L_0^{(\alpha)}(x) = 1$ e $L_1^{(\alpha)}(x) = 1+\alpha-x$. Existem várias relações que relacionam os polinómios de Laguerre para diferentes valores de α. Apenas indicamos as mais relevantes para os métodos espectrais (para mais detalhes ver [Sze39]),

$$\frac{d}{dx}L_{n+1}^{(\alpha)} = -L_n^{(\alpha)}, \quad n \in \mathbb{N}_0, \quad \alpha > -1, \tag{A.36}$$

$$L_{n+1}^{(\alpha)} = L_{n+1}^{(\alpha+1)} - L_n^{(\alpha+1)}, \quad n \in \mathbb{N}_0, \quad \alpha > -1. \tag{A.37}$$

A.4 Polinómios de Hermite

Os polinómios de Hermite, H_n, $n \in \mathbb{N}_0$ são funções próprias de um problema de Sturm Liouville não singular onde fazemos em (A.1) $p(x) = w(x) = e^{-x^2}$ e $q(x) = 0$, para todo $x \in I = \mathbb{R}$. Obtemos deste modo a equação diferencial

$$\left(e^{-x^2}H_n'(x)\right)' + 2ne^{-x^2}H_n(x) = 0 \tag{A.38}$$

A condição de normalização é

$$H_n(0) = (-1)^{n/2}\frac{n!}{(n/2)!} \quad \text{se } n \text{ é par} \tag{A.39}$$

$$H_n'(0) = (-1)^{(n-1)/2}\frac{(n+1)!}{((n+1)/2)!} \quad \text{se } n \text{ é impar} \tag{A.40}$$

A fórmula de Rodrigues toma a forma

$$H_n(x) = (-1)^n e^{x^2}\frac{d^n}{dx^n}e^{-x^2}, \tag{A.41}$$

explicitamente os polinómios de Hermite verificam a igualdade

$$H_n(x) = n!\sum_{m=0}^{[n/2]}(-1)^m\frac{(2x)^{n-2m}}{m!(n-2m)!}, \quad n \in \mathbb{N}_0, \quad x \in \mathbb{R}, \tag{A.42}$$

logo os polinómios de Hermite são funções pares (ímpares) se n é par (ímpar). A respectiva relação de recorrência a três termos é

$$H_n(x) = 2xH_{n-1}(x) - 2(n-1)H_{n-2}(x), \quad n \geq 2, \tag{A.43}$$

onde $H_0(x) = 1$ e $H_1(x) = 2x$. As derivadas dos polinómios de Hermite tomam uma forma particularmente simples. Derivando (A.41) temos que

$$H_n'(x) = 2xH_n(x) - H_{n+1}(x), \quad n \in \mathbb{N}, \quad x \in \mathbb{R}, \tag{A.44}$$

e usando a relação de recorrência (A.43) obtemos

$$H_n'(x) = 2nH_{n-1}(x), \quad n \geq 1, \quad x \in \mathbb{R}. \tag{A.45}$$

A.5 Ortogonalidade

O resultado fundamental desta secção é que verificadas certas condições as funções próprias de um problema de Sturm-Liouville formam um sistema ortogonal relativamente a um produto interno que depende da função w, frequentemente designada por *função peso*.

A.5.1 Polinómios Ortogonais

Assumindo que a função p se anula nos extremos do intervalo I, caso I seja limitado, ou que $\lim_{x\to\pm\infty} p(x) = 0$ caso I seja ilimitado. Então é válido o seguinte [Fun92]

Teorema A.5.1. Seja $\{u_n\}_{n\in\mathbb{N}_0}$ uma sucessão de funções próprias do problema (A.1) e $\{\lambda_n\}_{n\in\mathbb{N}_0}$ a sucessão de valores próprios correspondentes. Se a sucessão de valores próprios verificar a condição $\lambda_n \neq \lambda_m$ se $n \neq m$ então, tem-se

$$\int_I u_n u_m w dx = 0, \quad \forall\ n, m \in \mathbb{N}_0. \tag{A.46}$$

Deste modo, as famílias dos polinómios consideradas atrás nomeadamente, os polinómios de Jacobi, Laguerre e Hermite são ortogonais relativamente ao produto interno

$$(u, v)_w = \int_I uvw dx. \tag{A.47}$$

A norma, chamada de norma-w, associada ao produto interno (A.47) é definida por

$$||u||_w = \sqrt{(u, u)_w} = \sqrt{\int_I u^2 w dx}. \tag{A.48}$$

As normas dos polinómios de Jacobi são dadas por [Sze39],

$$\int_{-1}^{1}\left[P_n^{(\alpha,\beta)}\right]^2(1-x)^\alpha(1+x)^\beta dx=\begin{cases}2^{\alpha+\beta+1}\frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} & \text{se } n=0,\\ \\ \frac{2^{\alpha+\beta+1}}{(2n+\alpha+\beta+1)n!}\frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1)} & \text{se } n>0\end{cases}. \quad (A.49)$$

Em, particular para os polinómios de Legendre e de Chebyshev tem-se,

$$\int_{-1}^{1}P_n^2(x)dx=\frac{2}{2n+1},\quad n\in\mathbb{N}_0, \quad (A.50)$$

$$\int_{-1}^{1}T_n^2(x)\frac{dx}{\sqrt{1-x^2}}=\begin{cases}\pi & \text{se } n=0\\ \\ \frac{\pi}{2} & \text{se } n>0\end{cases}. \quad (A.51)$$

e para os polinómios de Laguerre e de Hermite,

$$\int_0^{+\infty}\left[L_n^\alpha(x)\right]^2x^\alpha e^{-x}dx=\frac{\Gamma(n+\alpha+1)}{n!},\quad n\in\mathbb{N}_0, \quad (A.52)$$

$$\int_{-\infty}^{+\infty}H_n^2(x)e^{-x^2}dx=2^n n!\sqrt{\pi}. \quad (A.53)$$

Seja $\{u_n\}_{n\geq 0}$ uma família de soluções polinomiais de um problema de Sturm-Liouville com função peso w e seja $\{\tilde{u}_n\}_{n\geq 0}$ família de polinómios mónicos associada. Ou seja, tem-se

$$\tilde{u}_n=\gamma_n^{-1}u_n,\quad \text{onde } \gamma_n=\lim_{x\to+\infty}\frac{u_n(x)}{x^n}, \quad (A.54)$$

então é válido o seguinte

Teorema A.5.2. Para todo o $n\in\mathbb{N}_0$ e para todo o polinómio, p, de grau n tal que $p(x)=x^n+\cdots$ tem-se

$$||\tilde{u}_n||_w\leq||p||_w. \quad (A.55)$$

A.5.2 Coeficientes de Fourier

Dado que os polinómios u_k, $k=0,1,\ldots,n$, são ortogonais relativamente a um produto interno, formam uma base do espaço vetorial, $\mathcal{P}_n$, dos polinómios de grau não superior a n. $\mathcal{P}_n$ tem dimensão $n+1$ logo, para todo $p\in\mathcal{P}_n$, podemos determinar univocamente $n+1$ coeficientes c_k, $k=0,1,\ldots,n$ tal que

$$p=\sum_{k=0}^{n}c_k u_k. \quad (A.56)$$

Os coeficientes c_k chamam-se *coeficientes de Fourier* de p relativamente à base considerada. Dados $p = \sum_{k=0}^{n} c_k u_k$ e $q = \sum_{k=0}^{n} b_k u_k$ então

$$pq = \sum_{k=0}^{n} \sum_{j=0}^{n} (c_k b_j) u_k u_j. \tag{A.57}$$

Integrando (A.57) em I e usando a ortogonalidade obtemos

$$(p, q)_w = \sum_{k=0}^{n} c_k b_k ||u_k||_w^2, \tag{A.58}$$

logo

$$||p||_w = \left(\sum_{k=0}^{n} c_k^2 ||u_k||_w^2 \right)^{1/2}. \tag{A.59}$$

Em particular se $q = u_k$, $k = 0, 1, \ldots, n$, temos a expressão explicita para os os coeficientes de Fourier de p

$$c_k = \frac{(p, u_k)_w}{||u_k||_w^2}, \quad p \in \mathcal{P}_n, \quad k = 0, 1, \ldots, n. \tag{A.60}$$

Seja f uma função tal que fw é integrável em I (caso I seja ilimitado exigimos também que f tenda para zero no infinito), então define-se os coeficientes de f fazendo

$$c_k = \frac{1}{||u_k||_w^2} \int_I f u_k w dx, \quad k \in \mathbb{N}_0. \tag{A.61}$$

Definimos o *operador projecção*

$$\Pi_{w,n} : C^0(\bar{I}) \longrightarrow \mathcal{P}_n, \quad n \in \mathbb{N}_0, \tag{A.62}$$

tal que $\Pi_{w,n} f = p_n$, com $p_n = \sum_{k=0}^{n} c_k u_k$. O operador projeção é linear e ao polinómio p_n chamamos projeção ortogonal de f em $\mathcal{P}_n$ relativamente ao produto interno $(*, *)_w$. O operador projeção satisfaz as propriedades seguintes

$$\Pi_{w,n} p = p, \quad \forall p \in \mathcal{P}_n, \tag{A.63}$$

$$\Pi_{w,n} u_m = 0, \quad \forall m > n. \tag{A.64}$$

A.6 Norma Infinito

Seja I um intervalo limitado. No espaço das funções contínuas em $\bar{I}$. Definimos a norma infinito

$$||f||_\infty = \max_{x \in \bar{I}} |f(x)| \quad \forall f \in C^0(\bar{I}). \tag{A.65}$$

Um dos principais resultados que caracterizam os polinómios de Chebyshev é o seguinte

Teorema A.6.1. Para todo $n \in \mathbb{N}_0$ e para todo o polinómio p em $\bar{I} = [-1, 1]$ de grau n com coeficiente guia unitário tem-se

$$||\tilde{T}||_\infty \leq ||p||_\infty \tag{A.66}$$

Note-se que

$$||\tilde{T}||_\infty = \begin{cases} 1 & \text{se } n = 0, \\ \\ 2^{1-n} & \text{se } n \geq 1. \end{cases}$$

Os dois teoremas, que se seguem, estabelecem limites superiores para a norma infinita da derivada de polinómios [Che66].

Teorema A.6.2 (Markoff)**.** Seja $\bar{I} = [-1, 1]$, então para todo $n \in \mathbb{N}_0$

$$||p'||_\infty \leq n^2 ||p||_\infty, \quad \forall p \in \mathcal{P}_n. \tag{A.67}$$

Teorema A.6.3 (Bernstein)**.** Seja $\bar{I} = [-1, 1]$, então para todo $n \in \mathbb{N}_0$

$$|p'(x)| \leq \frac{n}{\sqrt{1 - x^2}} ||p||_\infty, \quad \forall x \in I, \forall p \in \mathcal{P}_n. \tag{A.68}$$

Apêndice B

Espaços de Funções

Para o estudo das expansões de funções próprias de problemas de Sturm-Liouville é necessário especificar as classes a que pertencem as funções a aproximar. Estas classes são definidas usando o conceito de integral de de Lebesgue.

B.1 O Integral de Lebesgue

A integração de Lebesgue permite alargar a integração a classes de funções que não são integráveis no sentido de Riemann.

Começamos por enumerar os conceitos e as propriedades, relevantes para este trabalho, da teoria da medida de Lebesgue para o caso unidimensional.

Seja $I =]\,a, b\,[$, $a < b$ um intervalo limitado. A *medida de Lebesgue* μ de um subconjunto $J \subset \bar{I}$ é um número real não negativo. O conjunto vazio tem medida zero. Se $J \subset \bar{I}$ tem apenas um ponto ou um número finito de pontos, então $\mu(J) = 0$. Se J é um intervalo tal que $\bar{J} = [c, d]$, então $\mu(J) = d - c$. Quando J é a união de m intervalos disjuntos, $K_i,\ i = 1, 2, \cdots, m$, designados por *conjuntos elementares*, então $\mu(J) = \sum_{i=1}^{m} \mu(K_i)$. A união, intersecção e diferença de dois conjuntos elementares é ainda um conjunto elementar. Se J é a união de uma família contável de intervalos disjuntos, $K_i \subset \bar{I},\ i \in \mathbb{N}$ então a série $\sum_{i=1}^{\infty} \mu(K_i)$ de números não negativos é limitada por $b - a$ e consequentemente é convergente, sendo a sua soma a medida de Lebesgue do conjunto J, $\mu(J) = \sum_{i=1}^{\infty} \mu(K_i)$. No caso dos intervalos K_i não serem disjuntos é possível encontrar uma família $\tilde{K}_i$ de intervalos disjuntos tais que $\bigcup_{i\geq 1} \tilde{K}_i = J$ e tem-se

$$\mu(J) = \sum_{i=1}^{\infty} \mu(\tilde{K}_i) \leq \sum_{i=1}^{\infty} \mu(K_i). \tag{B.1}$$

Seja $J \subset \bar{I}$, define-se *medida externa* μ^* de J por

$$\mu^*(J) = \inf \left\{ \sum_{i=1}^{\infty} \mu(K_i) \mid J \subset \bigcup_{i\geq 1} \tilde{K}_i \right\}, \tag{B.2}$$

onde o ínfimo é tomado sobre todas as coberturas de J de famílias contáveis de intervalos K_i, $i \geq 1$. Para conjuntos elementares a medida μ^* coincide com μ. Deste modo dizemos que um conjunto $J \subset \bar{I}$ é *mensurável* se e somente se

$$\mu^*(J) = \mu(\bar{I}) - \mu^*(\bar{I} - J). \tag{B.3}$$

Para os conjuntos mensuráveis J tem-se $\mu^*(J) = \mu(J)$. A união e a intersecção de um conjunto finito de conjuntos mensuráveis é mensurável. Dado uma família J_i, $i \in \mathbb{N}$, de conjuntos mensuráveis e disjuntos dois a dois, então

$$\mu\left(\bigcup_{i\geq 1} J_i\right) = \sum_{i=1}^{\infty} \mu(J_i). \tag{B.4}$$

Uma função $f : \bar{I} \longrightarrow \mathbb{R}$ diz-se mensurável se o conjunto $\left\{x \in \bar{I} \,|f(x) < \gamma\right\}$ é mensurável para todo $\gamma \in \mathbb{R}$. A soma e o produto de funções mensuráveis é uma função mensurável. Se o conjunto $f(\bar{I})$ for finito ou contável dizemos que f é uma *função simples*. Toda a função limitada mensurável pode ser representada como um limite uniforme de funções simples f_n, $n \in \mathbb{N}$, ou seja

$$\lim_{n\to\infty} \sup_{x\in I} |f(x) - f_n(x)| = 0.$$

Duas funções mensuráveis f e g coincidem em quase todo o ponto (q.t.p.) quando $\mu\left(\left\{x \in \bar{I} \mid f(x) \neq g(x)\right\}\right) = 0$. Nesta situação dizemos que as funções f e g são *equivalentes*.

Seja f uma função simples. Então atinge, no máximo, um conjunto contável $\{\gamma_i\}$, $i \in \mathbb{N}$, e os conjuntos $A_i = \left\{x \in \bar{I} \mid f(x) = \gamma_i\right\}$, $i \in \mathbb{N}$ são mensuráveis. Se a série $\sum_{i=1}^{\infty} \gamma_i \mu\left(A_i\right)$ for convergente dizemos que f é integrável no sentido de Lebesque e definimos o integral de Lebesgue da função f no intervalo I, da forma

$$\int_I f dx = \sum_{i=1}^{\infty} \gamma_i \mu\left(A_i\right). \tag{B.5}$$

Dada uma função mensurável f, existe uma sucessão de funções simples f_n, $n \in \mathbb{N}$ que tende uniformemente para f. Uma função mensurável, f, é integrável no sentido de Lebesgue se existir $\lim_{n\to\infty} \int_I f_n dx$. Neste caso o limite não depende da sucessão aproximante $\{f_n\}_{n\in\mathbb{N}}$, e definimos o integral de Lebesgue de f como sendo

$$\int_I f dx = \lim_{n\to\infty} \int_I f_n dx. \tag{B.6}$$

A soma de duas funções integráveis é ainda uma função integrável e os integrais de Riemann de funções contínuas ou monótonas coincidem com os integrais de Lebesgue. Evidentemente que os integrais de duas funções mensuráveis equivalentes são iguais.

B.2 Espaços de Funções Mensuráveis

Seja f uma função mensurável, representamos por $[f]$ a classe de toda as funções equivalentes a f. Dada uma função peso $w : I \longrightarrow \mathbb{R}$ contínua e positiva definimos o seguinte espaço de funções

$$L^2_w(I) = \left\{ [f] \mid f \text{ é mensurável e } \int_I f^2 w dx < +\infty \right\}. \tag{B.7}$$

Definimos em $L^2_w(I)$ um produto interno e uma norma

$$(f,g)_{L^2_w(I)} = \int_I fgwdx, \quad \forall f,g \in L^2_w(I), \tag{B.8}$$

$$||f||_{L^2_w(I)} = \left(\int_I f^2 w dx \right)^{\frac{1}{2}}, \quad \forall f \in L^2_w(I). \tag{B.9}$$

Uma função, f, diz-se *essencialmente limitada* quando existe uma constante $M \geq 0$ tal que $|f(x)| \leq M$, em quase todo o ponto x .

Definimos

$$L^\infty(I) = \{[f] \mid f \text{ é essencialmente limitada}\} \tag{B.10}$$

e a correspondente norma

$$||f||_{L^\infty(I)} = \inf \{M \geq 0 \mid |f| \leq M \text{ q.t.p.}\}. \tag{B.11}$$

Quando I é limitado temos que, $C^0(\bar{I}) \subset L^\infty(I)$ e a norma infinito em $C^0(\bar{I})$ coincide com a norma (B.11).

Seja $\mathbf{X}$ um espaço vetorial normado, dizemos que uma sucessão, $\{x_n\}_{n\in\mathbb{N}}$ em $\mathbf{X}$, é de Cauchy quando para todo $\epsilon > 0$, podemos encontrar $n \in \mathbb{N}$ tal que

$$||x_{m_1} - x_{m_2}|| < \epsilon, \quad \forall m_1, m_2 > n. \tag{B.12}$$

Toda a sucessão convergente em $\mathbf{X}$ é de Cauchy. Dizemos que $\mathbf{X}$ é um *espaço normado completo* (ou espaço de Banach) se toda a sucessão de Cauchy for convergente em $\mathbf{X}$. Se I é um intervalo limitado então o espaço $C^0(\bar{I})$ com a norma $||*||_w$ definida em (A.48) não é completo, contudo o mesmo espaço equipado com a norma infinito, definida em (A.65), é um espaço completo e o espaço $L^\infty(I)$ equipado com a norma (B.11) é igualmente completo.

Um resultado importante é que $L^2_w(I)$ munido com a norma definida em (B.9) é um espaço completo. Quando o intervalo I é limitado $L^\infty(I)$ é o fecho de $L^2_w(I)$, ou seja, toda a função em $L^2_w(I)$ pode ser aproximada, relativamente à norma definida em (B.9) por uma sucessão de funções contínuas. Neste caso, além de termos $C^0(\bar{I}) \subset L^2_w(I)$, existe uma constante K que depende de $\mu(I)$ e de w tal que

$$||f||_{L^2_w(I)} \leq K||f||_{C^0(\bar{I})}, \quad \forall f \in C^0(\bar{I}). \tag{B.13}$$

Um espaço completo diz-se um espaço de Hilbert se a norma associada for definida por um produto interno. Deste modo $L^2_w(I)$ é um espaço de Hilbert e $C^0(\bar{I})$ não é um espaço de Hilbert.

B.3 Derivadas Fracas

Consideremos o espaço $C^k(\bar{I})$, onde $k \in \mathbb{N}$ das funções cuja derivadas de ordem inferior ou igual a k existem e são contínuas em $\bar{I}$. Quando I é limitado, definimos a norma

$$||f||_{C^k(\bar{I})} = ||f||_{C^0(\bar{I})} + \left\|\frac{df}{dx}\right\|_{C^0(\bar{I})} + \cdots + \left\|\frac{d^k f}{dx^k}\right\|_{C^0(\bar{I})} = \sum_{m=1}^{k} \left\|\frac{d^k f}{dx^k}\right\|_{C^0(\bar{I})}, \quad f \in C^k(\bar{I}),\ k \in \mathbb{N}. \tag{B.14}$$

$C^k(\bar{I})$ equipado com a norma (B.14) é um espaço completo. Definimos o espaço $C^\infty(\bar{I})$, das funções f tais que $f \in C^k(\bar{I})$, para todo $k \in \mathbb{N}$.

Consideremos o conjunto C_0^∞ definido por, $\phi \in C_0^\infty(I)$ se e somente se $\phi \in C^\infty(I)$ e existe um subconjunto próprio de I, fechado e limitado J_ϕ tal que ϕ se anula em $I \setminus J_\phi$. Iremos considerar, daqui para a frente, apenas subconjuntos de funções contínuas. Uma função, f, integrável no sentido de Lebesgue é derivável no sentido fraco se existe outra função g igualmente integrável tal que

$$\int_I f\phi' dx = -\int_I g\phi dx, \quad \forall \phi \in C_0^\infty(I). \tag{B.15}$$

A função g diz-se a *derivada fraca* de f e é indicada usando a notação da derivação convencional. Esta definição não entra em conflito com a derivação convencional. De facto se $f \in C^1(\bar{I})$ então $g \in C^0(\bar{I})$ e g coincide com a derivada convencional de f.

As derivadas de ordens superiores definem-se de modo análogo. A função mensurável, g é a $k-$ésima derivada de f se

$$\int_I f\frac{d^k\phi}{dx^k} dx = (-1)^k \int_I g\phi dx, \quad \forall \phi \in C_0^\infty(I). \tag{B.16}$$

A derivada fraca não é univocamente determinada, contudo duas derivadas fracas de uma função são equivalentes. Geralmente nem todas as funções possuem derivadas fracas, a não ser que se alargue o espaço das derivadas possíveis (ver p.e. [Sch66]).

B.4 Espaços de Sobolev pesados em intervalos

Dado $k \in \mathbb{N}$, define-se o espaço de Sobolev de ordem k num intervalo $I \subset \mathbb{R}$ relativamente ao peso w como sendo

$$H^k_w(I) = \left\{ [f] \ | \ f \text{ é } k \text{ vezes derivável e } \frac{d^m f}{dx^m} \in L^2_w(I), m = 0, 1, \cdots, k \right\}. \tag{B.17}$$

A função peso w é a função peso associada ao espaço $L^2_w(I)$, e para $k = 0$ tem-se $H^0_w(I) = L^2_w(I)$. Para todo o $k \in \mathbb{N}_0$, $H^k_w(I)$ é um espaço de Hilbert equipado com o produto interno e respetiva norma associada

$$(f,g)_{H^k_w(I)} = \sum_{m=0}^{k} \left(\frac{d^m f}{dx^m}, \frac{d^m f}{dx^m} \right)_{L^2_w(I)}, \quad \forall f, g \in H^k_w(I), \tag{B.18}$$

$$\|f\|_{H^k_w(I)} = \left(\sum_{m=0}^{k} \left\| \frac{d^m f}{dx^m} \right\|^2_{L^2_w(I)} \right)^{\frac{1}{2}}, \quad \forall f, g \in H^k_w(I). \tag{B.19}$$

Iremos listar apenas algumas propriedades destes espaços (para mais detalhes ver [Kuf85]). Uma propriedade fundamental é que para toda a função $f \in H^k_w(I)$ existe uma sucessão de funções regulares $\{g_n\}_{n\in\mathbb{N}}$, por exemplo $g_n \in C^\infty(\bar{I})$, $\forall n \in \mathbb{N}$, tal que $\lim_{n\to+\infty} \|f - g_n\|_{H^k_w(I)} = 0$. Muitos resultados são demonstrados usando esta propriedade, ou seja demonstram-se para as funções regulares e alarga-se, passando a limites, a espaços de Sobolev.

Consideremos as funções peso de Jacobi $w(x) = (1-x)^\alpha (1+x)^\beta$, $x \in I = [-1,1]$, onde os parâmetros satisfazem $-1 < \alpha, \beta < 1$. Verifica-se neste caso que: $H^1_w(I) \subset C^0(\bar{I})$ e existem duas constantes $K_1, K_2 > 0$ tais que

$$\|f\|_{L^2_w(I)} \leq K_1 \|f\|_{C^0(\bar{I})} \leq K_2 \|f\|_{H^1_w(I)}, \quad \forall f \in H^1_w(I). \tag{B.20}$$

Usando a dupla desigualdade (B.20) demonstra-se a relação

$$\left\| \frac{f}{1-x^2} \right\|_{L^2_w(I)} \leq K \|f'\|, \quad \forall f \in H^1_w(I) \text{ com } f(\pm 1) = 0. \tag{B.21}$$

Introduzindo o espaço

$$H^1_{0,w}(I) = \left\{ f \in H^1_w(I) \mid f(\pm 1) = 0 \right\}, \tag{B.22}$$

adoptamos em $H^1_{0,w}(I)$ o seguinte produto interno

$$[f,g]_w = \int_I \frac{df}{dx} \frac{dg}{dx} w dx \quad \forall f, g \in H^1_{0,w}(I), \tag{B.23}$$

e a norma induzida em $H^1_{0,w}(I)$

$$\|f\|_{H^1_{0,w}(I)} = \|f'\|_{L^2_w(I)}, \quad \forall f \in H^1_{0,w}(I), \tag{B.24}$$

que, devido à desigualdade de Poincaré (ver B.4.1) é equivalente à norma $\|*\|_{H^1_w(I)}$.

Análogamente define-se

$$P^0_N = \{\phi \in \mathcal{P}_N \mid \phi \text{ anula-se nos extremos de } I\}. \tag{B.25}$$

B.4.1 Desigualdade de Poincaré

Para domínios de dimensão $d = 1$ tem-se

Teorema B.4.1 (Desigualdade de Poincaré em espaços de Sobolev num intervalo I)**.** Seja $\bar{H}^1_w(I) = \left\{v \in H^1(I) \mid \exists x_0 \in \bar{I},\ v(x_0) = 0\right\}$. Então existe uma constante $C(I) > 0$ tal que

$$\|v\|_{L^2(I)} \leq C(I) \left\| \frac{dv}{dx} \right\|_{L^2(I)}, \quad \forall v \in \bar{H}^1_w(I) \tag{B.26}$$

Teorema B.4.2 (Desigualdade de Poincaré em espaços de Sobolev pesados (com peso w) num intervalo I)**.** Seja $\bar{H}^1_w(I) = \left\{v \in H^1_w(I) \mid \exists x_0 \in \bar{I},\ v(x_0) = 0\right\}$. Então existe uma constante $C(I) > 0$ tal que

$$\|v\|_{L^2_w(I)} \leq C(b-a) \left\| \frac{dv}{dx} \right\|_{L^2_w(I)}, \quad \forall v \in \bar{H}^1_w(I) \tag{B.27}$$

Para domínios de dimensão $d > 1$. Tem-se:

Teorema B.4.3 (Desigualdade de Poincaré em espaços de Sobolev num domínio Ω)**.** Seja $H^1_0(\Omega) = \{v \in H^1(\Omega) \mid \ v \mid_{\partial\Omega} = 0\}$. Existe uma constante $C(\Omega) > 0$ tal que

$$\|v\|_{L^2(\Omega)} \leq C(\Omega) \|\nabla v\|_{L^2(\Omega)}, \quad \forall v \in H^1_0(\Omega), \tag{B.28}$$

Teorema B.4.4 (Desigualdade de Poincaré em espaços de Sobolev pesados num domínio Ω)**.**
Seja $H^1_{w,0}(\Omega) = \left\{v \in H^1_{w,0}(\Omega) \mid \ v \mid_{\partial\Omega} = 0\right\}$. Existe uma constante $C(\Omega) > 0$ tal que

$$\|v\|_{L^2_w(\Omega)} \leq C(\Omega) \|\nabla v\|_{L^2_w(\Omega)}, \quad \forall v \in H^1_{0,w}(\Omega), \tag{B.29}$$

onde o operador *gradiente* é representado pelo símbolo ∇. As mesmas desigualdades são válidas se o domínio é simplesmente conexo e para o conjunto das funções v que se anulam num subconjunto de $\partial\Omega$ com medida positiva.

Apêndice C

Aproximação Polinomial

Iremos apenas mencionar os resultados mais relevantes para os métodos espectrais. Ao longo desta secção consideramos que C representa uma constante positiva que apenas depende da norma envolvida (não depende: das funções envolvidas, do diâmetro do domínio nem do inteiro positivo N). Iremos usar $P_N u$ para representar a expansão parcial de ordem N em funções próprias de problemas de Sturm-Liouville (S-L) singulares da função u e o *erro de truncatura de ordem* N é definido como sendo $u - P_N u$.

C.1 Expansões em Funções Próprias de Problemas S-L Singulares

Começamos com um resultado, que sob determinadas condições, garante que o espectro de um problema de S-L satisfaz as condições mencionadas em A.1. Ou seja os valores próprios $\{\lambda_n\}_{n>0}$ formam uma sucessão ilimitada de números reais positivos.

Consideramos no problema de S-L (A.1) com, $I = [-1, 1]$ e $p(-1) = p(1) = 0$ onde a solução u satisfaz a condição

$$\lim_{x \to \pm 1} p(x)u'(x) = 0.$$

Seja $X = \left\{ v \in L^2_w(I) \cap L^2_q(I) \mid \ v' \in L^2_p(I) \right\}$, X é um espaço de Hilbert, com a norma

$$\|v\| = \sqrt{\int_{-1}^{1} v^2 w dx + \int_{-1}^{1} v^2 p dx + \int_{-1}^{1} (v')^2 p dx}.$$

Assumindo que $u \in X$, é possível representar o problema (A.1) na formulação variacional

$$\int_{-1}^{1} (pu'v' + quv)dx = \lambda \int_{-1}^{1} uvw, \quad \forall v \in X. \tag{C.1}$$

dadas as condições acima tem-se [CHQZ07]

Teorema C.1.1. Os valores próprios da formulação variacional (C.1) formam uma sucessão de números reais $0 \leq \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_n \leq \cdots$. Além disso, para todo $k \in \mathbb{N}_0$ λ_k tem multiplicidade finita, e, o sistema $\{\phi_k\}_{k\in\mathbb{N}_0}$ formado pelas correspondentes funções próprias gera um subespaço denso e é ortogonal em $L^2_w(I)$.

Definindo a sucessão dos coeficientes de Fourier $\hat{u}_k = (u, \phi_k)_w$ (assumindo a normalização $\|\phi_k\|_{L^2_w(I)} = 1$), onde $u \in L^2_w(I)$ e usando a formulação variacional (com $v = \phi_k$) tem-se

$$\begin{aligned} \hat{u}_k &= \frac{1}{\lambda_k}\int_{-1}^{1}(p\phi'_k u' + q\phi_k u)\,dx \\ &= \frac{1}{\lambda_k}\int_{-1}^{1}(-(pu')' + qu)\phi_k dx + \frac{1}{\lambda_k}\left[pu'\phi_k\right]_{-1}^{1} \qquad \text{(C.2)} \\ &= \frac{1}{\lambda_k}\left(\frac{1}{w}\mathscr{L}u, \phi_k\right)_w + \left[pu'\phi_k\right]_{-1}^{1}. \end{aligned}$$

Esta igualdade é válida se a função

$$u^{(1)} = \frac{1}{w}\mathscr{L}u \in L^2_w(I). \qquad \text{(C.3)}$$

Sob esta hipótese, pu' é contínua em I uma vez que se tem

$$|(pu')(x_1) - (pu')(x_2)| = \left|\int_{x_1}^{x_2}(pu')'dx\right| \leq \left(\int_{x_1}^{x_2}\frac{1}{w}|(pu')'|^2\right)^{1/2}\left(\int_{x_1}^{x_2}w\right)^{1/2}.$$

Devido à regularidade do operador elíptico $\mathscr{L}$, temos que a $u'' \in \mathscr{L}^2_{1/w}(I)$. Então u e u' são contínuas em I e devido às condições fronteira (C.1), tem-se que $\left[pu'\phi_k\right]_{-1}^{1} = 0$. Deste modo

$$\hat{u}_k = \frac{1}{\lambda_k}\left(u^{(1)}, \phi_k\right)_w.$$

Se $u_{(m)} = \frac{1}{w}\mathscr{L}u^{(m-1)} \in L^2_w(I)$ e $u^{(m-1)}$ satisfaz as condições fronteira (C.1) para $m \geq 2$ então,

$$\hat{u}_k = \frac{1}{(\lambda_k)^m}\left(u^{(m-1)}, \phi_k u\right). \qquad \text{(C.4)}$$

No caso dos polinómios de Legendre ou de Chebyshev tem-se respetivamente $\lambda_k = k(k+1)$ e $\lambda_k = k^2$ (ver (A.2.1) e (A.2.2)) logo o decaimento dos coeficientes de Fourier de uma função $u \in C^\infty(I)$ é mais rápido do que qualquer potência negativa de k. Iremos de seguida apresentar os resultados mais relevantes para aproximações expandidas em polinómios de Legendre e de Chebyshev.

C.2 Aproximações de Legendre

Os resultados relativos às desigualdades inversas a respeito da somabilidade e derivação de polinómios algébricos são expressos em termos de normas em espaços de Banach $L^p(I)$, $1 \leq p < \infty$ e $L^\infty(I)$.

Teorema C.2.1. (ver [Tim63]) Seja $I =]a, b[$ então para todos p, q tais que $1 \leq p \leq q \leq \infty$ existe uma constante positiva C[1] independente de N tal que

$$\|\phi\|_{L^q(I)} \leq C N^{2(1/p-1/q)} \|\phi\|_{L^p(I)}, \quad \forall \phi \in \mathcal{P}_N. \tag{C.5}$$

Considerando a função $\eta_\alpha = (1-x^2)^\alpha$, $\alpha \geq 0$ tem-se

Teorema C.2.2. (ver Bernardi e Maday (1997a))) Existe uma constante $C > 0$ independente de N tal que

$$\|\phi\|_{L^2(I)} \leq C N^\alpha \|\phi\|_{L^2_{\eta_\alpha}(I)}, \quad \forall \phi \in \mathcal{P}_N. \tag{C.6}$$

Relativamente à derivação tem-se o seguinte resultado geral (ver [Tim63] para $p = \infty$; para $p = 2$ ver Babuska(1981) e para $2 < p < \infty$ ver Quarteroni (1984))

Teorema C.2.3. Para todo p, tal que $2 \leq p \leq \infty$ e para todo $k \in \mathbb{N}$ existe uma constante positiva C independente de N tal que

$$\left\| \frac{d^k \phi}{dx^k} \right\|_{L^p(I)} \leq C N^{2k} \|\phi\|_{L^p(I)}, \quad \forall \phi \in \mathcal{P}_N. \tag{C.7}$$

Fazendo em (C.6), $\eta(x) = (1-x^2)$, obtemos as desigualdades (ver Bernardi e Maday (1997a)))

$$\left\| \sqrt{\eta} \frac{d\phi}{dx} \right\|_{L^2(I)} \leq \sqrt{2} N \|\phi\|_{L^2(I)}, \quad \forall \phi \in \mathcal{P}_N, \tag{C.8}$$

e caso o polinómio ϕ se anule nos extremos de I tem-se

$$\left\| \frac{d\phi}{dx} \right\|_{L^2(I)} \leq \sqrt{2} N \left\| \frac{\phi}{\sqrt{\eta}} \right\|_{L^2(I)}, \quad \forall \phi \in \mathcal{P}_N^0(I). \tag{C.9}$$

Outra desigualdade que permite limitar a norma infinito de um polinómio pela sua norma num espaço de Sobolev de ordem 1/2 (`acrescentar espaços de Sobolev de ordens fraccionais`) é a seguinte (ver [CHQZ07])

$$\|\phi\|_{L^\infty(I)} \leq C \sqrt{\log(N+1)} \|\phi\|_{H^{1/2}(I)} \quad \forall \phi \in \mathcal{P}_N \tag{C.10}$$

[1] Mais precisamente tem-se $C = \left(\frac{2(p+1)}{b-a} \right)^{(1/p-1/q)}$

Apêndice D

Integração Numérica

D.1 Zeros de Polinómios Ortogonais

As fórmulas de quadratura de Gauss baseiam-se no conhecimento dos zeros de polinómios. Iremos caracterizar as suas propriedades de interesse para a teoria espectral, para mais detalhes ver p. ex. [Sze39].

Um resultado geral é o seguinte

Teorema D.1.1. Seja $\{u_n\}_{n\in\mathbb{N}_0}$ uma sucessão de soluções do problema de Sturm-Liouville,L (A.1), onde u_n é um polinómio de grau n que satisfaz a condição de ortogonalidade (A.46). Então, para todo $n \geq 1$, u_n tem exatamente n zeros reais e distintos em I.

Obviamente que o polinómio u_n' tem exatamente $n-1$ zeros reais e distintos em I. Iremos representar os n zeros de u_n por $\xi_k^{(n)}$, $1 \leq k \leq n$, onde assumimos que estão colocados por ordem crescente, $\xi_1^{(n)} < \xi_2^{(n)} < \ldots < \xi_n^{(n)}$, e os $n-1$ zeros de u_n' por $\eta_k^{(n)}$, $1 \leq k \leq n-1$.

Note-se que os polinómios u_n e u_n' não possuem zeros comuns assim como, entre dois zeros consecutivos de u_{n-1} existe um zero de u_n.

Um resultado igualmente relevante é o seguinte

Teorema D.1.2. Seja $\{u_n\}_{n\in\mathbb{N}_0}$ uma sucessão de polinómios ortogonais no intervalo I. Então, para todo o intervalo $[a,b] \subset I$ existe $m \in \mathbb{N}$ tal que u_m tem um zero em $[a,b]$.

Este resultado estabelece que o conjunto $\bigcup_{n\geq 1}\bigcup_{k=1}^{n}\left\{\xi_k^{(n)}\right\}$ é denso no intervalo $\overline{I}$.

Considerando os polinómios de Jacobi para os quais os parâmetros α e β satisfazem a condição $-\frac{1}{2} \leq \alpha, \beta \leq \frac{1}{2}$. Então, tem-se a seguinte estimativa para a localização dos zeros de $P^{(\alpha,\beta)_n}$, $n \geq 1$,

$$-1 \leq -\cos\frac{k+(\alpha+\beta-1)/2}{n+(\alpha+\beta+1)/2}\pi \leq \xi_k^{(n)} \leq -\cos\frac{k}{n+(\alpha+\beta+1)/2}\pi \leq 1, \quad 1 \leq k \leq n.$$

No caso dos polinómios ultra-esféricos $\left(-\frac{1}{2} \leq \alpha = \beta \leq \frac{1}{2}\right)$ é possível melhorar a estimativa anterior, [Fun92]

$$-1 \leq -\cos\frac{k+\frac{\alpha}{2}-\frac{1}{4}}{n+\alpha+\frac{1}{2}}\pi \leq \xi_k^{(n)} \leq -\cos\frac{k}{n+\alpha+\frac{1}{2}} \leq 0, \quad 1 \leq k \leq \left[\frac{1}{2}\right], \tag{D.1}$$

$$0 \leq \cos\frac{n-k+1}{n+\alpha+\frac{1}{2}}\pi \leq \xi_k^{(n)} \leq \cos\frac{n-k+\frac{\alpha}{2}+\frac{3}{4}}{n+\alpha+\frac{1}{2}}\pi, \quad n+1-\left[\frac{1}{2}\right] \leq k \leq n. \tag{D.2}$$

Para valores de $|\alpha| = |\beta| = \frac{1}{2}$ os zeros $\xi_k^{(n)}$, $1 \leq k \leq n$, são dados de forma exacta pelas expressões:

$$\xi_k^{(n)} = -\cos\frac{2k-1}{2n}\pi, \quad \text{para } \alpha = \beta = -\frac{1}{2}, \tag{D.3}$$

$$\xi_k^{(n)} = -\cos\frac{k}{n+1}\pi, \quad \text{para } \alpha = \beta = \frac{1}{2}, \tag{D.4}$$

$$\xi_k^{(n)} = -\cos\frac{2k}{2n+1}\pi, \quad \text{para } \alpha = \frac{1}{2},\ \beta = -\frac{1}{2}, \tag{D.5}$$

$$\xi_k^{(n)} = -\cos\frac{2k-1}{2n+1}\pi, \quad \text{para } \alpha = -\frac{1}{2},\ \beta = \frac{1}{2}. \tag{D.6}$$

Argumentos semelhantes são válidos para os zeros $\eta_k^{(n)}$ das derivadas dos polinómios ortogonais. Contudo será conveniente considerarmos os pontos extremos do intervalo $I = [-1, 1]$. Nestes casos, as seguintes igualdades:

$$\eta_k^{(n)} = -\cos\frac{k\pi}{n}, \quad \text{para } 0 \leq k \leq n \ \text{ e } \ \alpha = \beta = -\frac{1}{2}, \tag{D.7}$$

$$\eta_k^{(n)} = -\cos\frac{2k+1}{2n+2}\pi, \quad \text{para } 1 \leq k \leq n-1 \ \text{ e } \ \alpha = \beta = \frac{1}{2}, \tag{D.8}$$

$$\eta_k^{(n)} = -\cos\frac{2k+1}{2n+1}\pi, \quad \text{para } 1 \leq k \leq n \ \text{ e } \ \alpha = \frac{1}{2},\ \beta = -\frac{1}{2}, \tag{D.9}$$

$$\eta_k^{(n)} = -\cos\frac{2k}{2n+1}\pi, \quad \text{para } 0 \leq k \leq n-1 \ \text{ e } \ \alpha = -\frac{1}{2},\ \beta = \frac{1}{2}. \tag{D.10}$$

D.2 Bases de Lagrange

Um conjunto de polinómios ortogonais no intervalo I, $\{u_k\}_{0\leq k\leq n}$, é uma base de $\mathcal{P}_n$. Escolhidos $n+1$ pontos distintos em $\overline{I}$, $\{x_k\}_{0\leq k\leq n}$, existe uma base de $\mathcal{P}_n$ designada de *base de Lagrange* relativamente aos pontos $\{x_i\}_{0\leq i\leq n}$. A base de Lagrange, $\left\{l_j^{(n)}\right\}_{0\leq j\leq n}$ é univocamente determinada pelas condições

$$l_j^{(n)}(x_i) = \delta_{i,j}, \quad 0 \leq j \leq n, \tag{D.11}$$

e os elementos da base de Lagrange, chamados de *polinómios característicos de Lagrange* e são determinados por

$$l_j^{(n)}(x) = \prod_{\substack{i=1 \\ i\neq k}}^{n} \frac{x - x_i}{x_k - x_i}, \quad 0 \leq j \leq n. \tag{D.12}$$

Para todo $p \in \mathcal{P}_n$ tem-se

$$p = \sum_{j=0}^{n} p(x_j) l_j^{(n)} \tag{D.13}$$

Escolhidos para pontos os zeros $\xi_k^{(n)}$, $1 \leq k \leq n$, do polinómio de Jacobi $P_n^{(\alpha,\beta)}$, obtém-se uma base de Lagrange, $\left\{l_j^{(n)}\right\}_{1\leq j\leq n}$ para $\mathcal{P}_{n-1}$. Cada elemento desta base pode escrever-se da forma

$$l_j^{(n)}(x) = \begin{cases} \frac{u_n(x)}{u_n'\left(\xi_j^{(n)}\right)\left(x-\xi_j^{(n)}\right)} & \text{se } x \neq \xi_j^{(n)} \\ \\ 1 & \text{se } x = \xi_j^{(n)}, \end{cases} \tag{D.14}$$

onde $1 \leq j \leq n$ e $u_n = P_n^{(\alpha,\beta)}$, além disso, tem-se

$$\lim_{x\to\xi_j^{(n)}} l_j^{(n)}(x) = \lim_{x\to\xi_j^{(n)}} \frac{u_n'(x)}{u_n'(\xi_j^{(n)})} = 1, \quad 1 \leq j \leq n. \tag{D.15}$$

Por outro lado fixando os pontos $\eta_j^{(n)}$, $0 \leq j \leq n$, pode-se encontrar uma base de Lagrange, representada por $\tilde{l}_j^{(n)}$, de $\mathcal{P}_n$. Claro que tem-se

$$\tilde{l}_j^{(n)}\left(\eta_i^{(n)}\right) = \delta_{i,j}, \quad 0 \leq i, j \leq n, \tag{D.16}$$

e para todo $p \in \mathcal{P}_n$ tem-se

$$p = \sum_{j=0}^{n} p\left(\eta_j^{(n)}\right) \tilde{l}_j^{(n)}. \tag{D.17}$$

Analogamente a (D.14) tem-se o seguinte resultado para polinómios de Chebyshev, para o caso geral dos polinómios de Jacobi ver [Fun92].

Teorema D.2.1. Para todo $n \geq 0$,

$$\tilde{l}_j^{(n)} = \begin{cases} \frac{(-1)^n}{2n^2}(x-1)T_n'(x) & \text{se } j = 0, \\ \\ \frac{(-1)^{j+n}}{n^2}\frac{(x^2-1)T_n'(x)}{\left(x-\eta_j^{(n)}\right)} & \text{se } 1 \leq j \leq n-1, \\ \\ \frac{1}{2n^2}(x+1)T_n'(x) & \text{se } j = 1. \end{cases} \tag{D.18}$$

Seja uma família de polinómios ortogonais $\{u_n\}_{n\in\mathbb{N}_0}$ relativamente a uma função peso w num intervalo I. Dado $n \geq 1$, sejam $\xi_j^{(n)}$, $1 \leq j \leq n$ os zeros de u_n. Definimos o operador $I_{w,n} : C^0(I) \longrightarrow \mathcal{P}_{n-1}$ com sendo o operador que envia uma função contínua f no (único) polinómio p_n que satisfaz $p_n\left(\xi_j^{(n)}\right) = f\left(\xi_j^{(n)}\right)$, $1 \leq j \leq n$. O operador $I_{w,n}$ designa-se por *operador interpolador* e é um operador linear, além disso, tem-se

$$I_{w,n}p = p, \quad \forall p \in \mathcal{P}_{n-1}. \tag{D.19}$$

Usando os zeros da derivada e os pontos extremos do intervalo $]-1,1[$ definimos o operador interpolador $\tilde{I}_{w,n} : C^0(]-1,1[) \longrightarrow \mathcal{P}_n$, $n \geq 1$ tal que $\tilde{I}_{w,n}f$ é o (único) polinómio $p_n \in \mathcal{P}_n$ que verifica $p_n\left(\eta_j^{(n)}\right) = f\left(\eta_j^{(n)}\right)$, $0 \leq j \leq n$. $\tilde{I}_{w,n}$ também é um operador linear e tem-se

$$\tilde{I}_{w,n}p = p, \quad \forall p \in \mathcal{P}_n. \tag{D.20}$$

D.3 Fórmulas de integração de Gauss

Seja $\{u_k\}_{k\in\mathbb{N}_0}$ uma família de polinómios ortogonais num intervalo I relativamente a uma função peso w. Sejam $\xi_j^{(n)}$, $1 \leq j \leq n$ as raizes de u_n, usando (D.13) tem-se, para todo o $p \in \mathcal{P}_{n-1}$ verifica-se

$$\int_I pwdx = \sum_{j=1}^{n} p\left(\xi_j^{(n)}\right) w_j^{(n)}, \tag{D.21}$$

onde

$$w_j^{(n)} = \int_I l_j^{(n)} wdx \tag{D.22}$$

são os pesos da fórmula de integração gaussiana. A igualdade (D.21) é conhecida por *fórmula de quadratura de Gauss* e aos zeros de u_n, $\xi_j^{(n)}$, $1 \leq j \leq n$ chamam-se *nós da integração de Gauss*. O uso dos nós , $\xi_j^{(n)}$, $1 \leq j \leq n$ permite que a fórmula de quadratura de Gauss seja exacta para todo o polinómio de grau não superior a $2n-1$ (ver, p.e. [Fun92]). Os pesos (D.22) podem calcular-se de forma explícita (ver, p.e. [DR84]). Para os casos: de Legendre, de Chebyshev, de Laguerre e de Hermite tem-se:

- **Legendre:**
$$w_j^{(n)} = \frac{2}{n}\left(P_n\left(\xi_j^{(n)}\right) P_n'\left(\xi_j^{(n)}\right)\right)^{-1}, \quad 1 \leq j \leq n; \tag{D.23}$$

- **Chebyshev:**
$$w_j^{(n)} = \frac{\pi}{n}, \quad 1 \leq j \leq n; \tag{D.24}$$

- **Laguerre:** $(\alpha > -1)$
$$w_j^{(n)} = -\frac{\Gamma(n+\alpha)}{n!}\left[L_{n-1}^{(\alpha)}\left(\xi_j^{(n)}\right)\frac{d}{dx}L_n^{(\alpha)}\left(\xi_j^{(n)}\right)\right]^{-1}, \quad 1 \leq j \leq n; \tag{D.25}$$

- **Hermite:**

$$w_j^{(n)} = \sqrt{\pi} 2^{n+1} n! \left[H_n' \left(\xi_j^{(n)} \right) \right]^{-2}, \quad 1 \leq j \leq n. \tag{D.26}$$

D.4 Fórmulas de integração de Gauss-Lobato

As fórmulas de integração de Gauss-Lobato baseiam-se nos nós obtidos pelos zeros das derivadas de polinómios de Jacobi e pelos nós $\eta_0^{(n)} = -1$ e $\eta_n^{(n)} = 1$. Calculando $\int_I pwdx$, onde $p \in \mathcal{P}_n$ e usando (D.17) tem-se a *fórmula de integração de Gauss-Lobato*

$$\int_{-1}^{1} pwdx = \sum_{j=0}^{n} p\left(\eta_j^{(n)}\right) \tilde{w}_j^{(n)}, \tag{D.27}$$

onde w é a função peso de Jacobi e

$$\tilde{w}_j^{(n)} = \int_{-1}^{1} \tilde{l}_j^{(n)} wdx, \quad 0 \leq j \leq n, \tag{D.28}$$

são os *pesos de Gauss-Lobato*. Analogamente às fórmulas de integração de Gauss a integração de Gauss-Lobato é válida para polinómios de grau não superior a $2n-1$ (ver, p.e. [Fun92]). Os pesos de Gauss-Lobato para os casos de Legendre ou de Chebyshev podem determinar-se usando as seguintes igualdades:

- **Legendre:**

$$\tilde{w}_j^{(n)} = \begin{cases} \frac{2}{n(n+1)} & \text{se } j = 0 \text{ ou } j = 1, \\ \\ \frac{-2}{n+1} \left(P_n\left(\eta_j^{(n)}\right) P_n'\left(\eta_j^{(n)}\right) \right)^{-1} & \text{se } 1 \leq j \leq n-1; \end{cases} \tag{D.29}$$

- **Chebyshev:**

$$\tilde{w}_j^{(n)} = \begin{cases} \frac{\pi}{2n} & \text{se } j = 0 \text{ ou } j = 1, \\ \\ \frac{\pi}{n} & \text{se } \quad 1 \leq j \leq n-1. \end{cases} \tag{D.30}$$

Para os restantes polinómios de Jacobi e para mais detalhes ver p. ex. [DR84].

D.5 Normas discretas

Dois polinómios $p, q \in \mathcal{P}_{n-1}$ são univocamente determinados pelos valores que tomam nos nós de integração de Gauss, $\xi_j^{(n)}$, $1 \leq j \leq n$. Logo o produto interno $\int_I pqwdx$ pode-se determinar usando (D.21) dado que, o polinómio $pq \in \mathcal{P}_{2n-2}$. Contudo quando dois polinómios, p, q, são determinados pelos nós de Gauss-Lobato, $\eta_j^{(n)}$, $0 \leq j \leq n$ a fórmula

de integração de Gauss-Lobato (D.27) deixa de ser válida (dado que $pq \in \mathcal{P}_{2n}$), logo não podemos usar (D.27) para calcular o produto interno usual. Para ultrapassar esta dificuldade usa-se o chamado *produto interno discreto* e correspondente *norma discreta* associada definidos respetivamente por

$$(p,q)_{w,n} = \sum_{j=0}^{n} p\left(\eta_j^{(n)}\right) q\left(\eta_j^{(n)}\right) \tilde{w}_j^{(n)}, \quad \forall p,q \in \mathcal{P}_n, \tag{D.31}$$

e

$$\|p\|_{w,n} = \left(\sum_{j=0}^{n} p^2\left(\eta_j^{(n)}\right) \tilde{w}_j^{(n)}\right)^{1/2}, \quad \forall p,q \in \mathcal{P}_n. \tag{D.32}$$

D.6 Transformadas Discretas de Fourier

Todo o polinómio admite duas representações. Uma é dada pelos coeficientes de Fourier relativamente a um conjunto de funções base ortogonais, enquanto a outra é dada pelos valores que o polinómio toma num conjunto de nós associados com uma fórmula de integração Gaussiana. As seguintes transformações permitirão passar de uma representação para outra.

A igualdade (D.21) implica que para todo $n \geq 1$ tem-se,

$$\int_I l_i^{(n)} l_j^{(n)} w dx = \sum_{k=1}^{n} l_i^{(n)}\left(\xi_k^{(n)}\right) l_j^{(n)}\left(\xi_k^{(n)}\right) w_k^{(n)} = \delta_{i,j} w_i^{(n)}. \tag{D.33}$$

Para os nós de Gauss-Radau existe igualmente uma igualdade semelhante que garante a ortogonalidade dos polinómios da Lagrange. Para os nós de Gauss-Lobato os polinómios de Lagrange $\tilde{l}_i^{(n)}$, $0 \leq i \leq n$ deixam de ser ortogonais relativamente ao produto interno (A.47) dado que se tem [Fun92]

$$\int_{-1}^{1} \tilde{l}_i^{(n)} \tilde{l}_j^{(n)} w dx = \delta_{i,j} \tilde{w}_i^{(n)} - \frac{||u_n||_{w,n}^2 - ||u_n||_w^2}{||u_n||_{w,n}^4} u_n\left(\eta_i^{(n)}\right) u_n\left(\eta_j^{(n)}\right) \tilde{w}_i^{(n)} \tilde{w}_j^{(n)}, \tag{D.34}$$

para todo $0 \leq i,j \leq n$, onde $u_n = P_n^{(\alpha,\beta)}$, $\alpha, \beta \geq -1$.

Contudo os polinómios de Lagrange $\tilde{l}_i^{(n)}$, $0 \leq i \leq n$ são ortogonais relativamente ao produto interno (D.31), de facto tem-se

$$\left(\tilde{l}_i^{(n)}, \tilde{l}_j^{(n)}\right)_{w,n} = \delta_{i,j} \tilde{w}_i^{(n)}, \quad 0 \leq i,j \leq n. \tag{D.35}$$

Então para todo $n \geq 1$ existem duas bases ortogonais em $\mathcal{P}_{n-1}$, a base $\{u_k\}_{0 \leq k \leq n-1}$ e a base dos polinómios de Lagrange $\left\{l_j^{(n)}\right\}_{1 \leq j \leq n}$. Então podemos definir um automorfismo $K_n : \mathcal{P}_{n-1} \longrightarrow \mathcal{P}_{n-1}$, que envia as coordenadas de um polinómio $p \in \mathcal{P}_{n-1}$ na base

$\left\{l_j^{(n)}\right\}_{1\leq j\leq n}$ nas coordenadas de p na base $\{u_k\}_{0\leq k\leq n-1}$. A transformação K_n chama-se *transformada de Fourier discreta* e pode ser representada por uma matriz $\mathbf{K_n} \in \mathcal{M}_{n\times n}$. De facto, para cada coeficiente de Fourier $c_i = \frac{(p,u_i)_w}{||u_i||_w^2}$ tem-se,

$$c_i = \frac{1}{||u_i||_w^2} \sum_{j=1}^{n} p\left(\xi_j^{(n)}\right) u_i\left(\xi_j^{(n)}\right) w_j^{(n)}, \quad 0 \leq i \leq n-1. \tag{D.36}$$

Logo a matriz $\mathbf{K_n}$ tem entradas

$$\mathbf{K_n} = [k_{i,j}], \quad \text{onde} \quad k_{i,j} = \frac{u_i\left(\xi_{j+1}^{(n)}\right) w_{j+1}^{(n)}}{||u_i||_w^2}, \quad 0 \leq i,j \leq n-1. \tag{D.37}$$

Por outro lado, como para todo $p \in \mathcal{P}_{n-1}$ tem-se $p = \sum_{i=0}^{n-1} c_i u_i$, é possível determinar ${\mathbf{K_n}}^{-1}$ dado que

$$p\left(\xi_i^{(n)}\right) = \sum_{j=0}^{n-1} c_j u_j\left(\xi_i^{(n)}\right), \quad 1 \leq i \leq n, \tag{D.38}$$

logo

$$\mathbf{K_n^{-1}} = \left[u_j\left(\xi_{i+1}^{(n)}\right)\right], \quad 0 \leq i,j \leq n-1. \tag{D.39}$$

Após uma normalização apropriada dos elementos das duas bases conclui-se que a inversa de $\mathbf{K_n}$ é dada pela sua transposta dado que se tem $\mathbf{K_n^{-1}} = \mathbf{D_n}{\mathbf{K_n}}^T\overline{\mathbf{D_n}}$, onde $\mathbf{D_n}$ e $\overline{\mathbf{D_n}}$ são duas matrizes diagonais definidas por

$$\mathbf{D_n} = \operatorname{diag}\left[||u_j||_w^2\right]_{0\leq j\leq n-1} \quad \text{e} \quad \overline{\mathbf{D_n}} = \operatorname{diag}\left[\frac{1}{w_{i+1}^{(n)}}\right]_{0\leq i\leq n-1}.$$

Para distinguir a duas representações de um polinómio $p\mathcal{P}_{n-1}$, é frequentemente usada a seguinte terminologia. Quando o polinómio p é representado na base $\{u_k\}_{0\leq k\leq n-1}$ então $\mathcal{P}_{n-1}$ é isomorfo ao conjunto dos seus coeficientes de Fourier e diz-se que p está (representado) no *espaço das frequências*. No caso de p estar representado na base $\left\{l_j^{(n)}\right\}_{1\leq j\leq n}$ então $\mathcal{P}_{n-1}$ é isomorfo ao conjunto dos valores que os polinómios tomam nos nós e diz-se que p está (representado) no *espaço físico*.

Relativamente ao nós de Gauss-Lobato, temos que o espaço $\mathcal{P}_n$ possui duas base ortogonais $\{u_k\}_{0\leq k\leq n}$ e $\left\{\tilde{l}_j^{(n)}\right\}_{0\leq j\leq n}$. Neste caso a transformada discreta de Fourier $\tilde{K}_n : \mathcal{P}_n :\longrightarrow \mathcal{P}_n$ é representada por uma matriz $\tilde{\mathbf{K}}_\mathbf{n} \in \mathcal{M}_{(n+1)\times(n+1)}$. Os coeficientes de Fourier c_i, $0 \leq i \leq n$ são dados por (para mais detalhes ver [Fun92])

$$c_i = \begin{cases} \frac{1}{||u_i||_w^2} \sum_{j=0}^{n} p\left(\eta_j^{(n)}\right) u_i\left(\eta_j^{(n)}\right) \tilde{w}_j^{(n)} & \text{se } 0 \leq i \leq n-1, \\ \\ \frac{1}{||u_i||_{w,n}^2} \sum_{j=0}^{n} p\left(\eta_j^{(n)}\right) u_n\left(\eta_j^{(n)}\right) \tilde{w}_j^{(n)} & \text{se } i = n. \end{cases} \tag{D.40}$$

E, inversamente tem-se

$$p\left(\eta_j^{(n)}\right)\sum_{j=0}^{n} c_j u_j\left(\eta_i^{(n)}\right). \tag{D.41}$$

Observação: *A implementação da transformada discreta de Fourier tem um custo proporcional a n^2. Em especial quando se usa polinómios de Chebyshev é possível usar algoritmos mais eficientes para efectuar a transformada discreta de Fourier e a sua inversa. Estes algoritmos são conhecidos por transformadas rápidas de Fourier, para mais detalhes ver p.e [CHQZ07] ou [Fun92].*

D.6.1 *Aliasing*

Os operadores projecção (A.62) e os operadores de interpolação (D.19) e (D.20) não são iguais. Mais precisamente se $f \in C^0([-1,1])$ tem-se $\Pi_{w,n-1}f = I_{w,n-1}f$ sse $f \in \mathcal{P}_{n-1}$. Logo se $f \notin \mathcal{P}_{n-1}$ a passagem do espaço físico para o espaço das frequências e vice-versa causa um erro o chamado *efeito de aliasing*. Para cada $n \leq 1$ definimos o operador $A_{w,n} = I_{w,n} - \Pi_{w,n-1}$, e ao polinómio

$$A_{w,n}f = I_{w,n}f - \Pi_{w,n-1}f, \quad f \in C^0([-1,1]), \tag{D.42}$$

chama-se de *erro de aliasing*. Dependendo da regularidade da função f o erro de f tende para zero quando $n \to \infty$. Analogamente, para os nós de Gauss-Lobato definimos o erro de distorção como sendo o polinómio

$$\tilde{A}_{w,n}f = \tilde{I}_{w,n}f - \Pi_{w,n}f, \quad f \in C^0([-1,1]). \tag{D.43}$$

Bibliografia

[Ada78] R. A. Adams. *Sobolev spaces*. Pure and applied mathematics. Academic Press, New York, 1978.

[AS65] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1965.

[Ask75] R. Askey. *Orthogonal polynomials and special functions*. Regional Conference Series in Applied Mathematics. SIAM, 1975. (esp. lecture 7).

[BA72] I. Babušca and A.K. Aziz. Survey lectures on the mathematical foundations of the finite elements method. In A.K. Aziz, editor, *The Mathematical Foundations of the Finite Elements Method with Application to Partial Differential Equations*. Academic Press, New York, 1972.

[BGM96] George A. Baker and P. R. Graves-Morris. *Padé approximants*. Encyclopedia of Mathematics and its Applications, v. 59. Cambridge University Press, January 1996.

[BJ17] E. Borel and G. Julia. *Leçons sur les fonctions monogènes uniformes d'une variable complexe*. Gauthier-Villars, 1917.

[BJ73] George A. Baker Jr. The existence and convergence of subsequences of padé approximants. *J. Math. Comp. App.*, 43(2):498 – 528, 1973.

[BM14] Bernhard Beckermann and Ana C. Matos. Algebraic properties of robust padé approximants. *Journal of Approximation Theory*, (0), 2014.

[BMW08] Bernhard Beckermann, Ana C. Matos, and Franck Wielonsky. Reduction of the Gibbs phenomenon for smooth functions with jumps by the ϵ-algorithm. *J. Math. Comp. App.*, 219(2):329 – 349, 2008.

[Boy01] J. P. Boyd. *Chebyshev and Fourier Spectral Methods*. Dover Publications, Inc., 2001.

[Bre83] H. Brezis. *Analyse fonctionnelle : théorie et applications*. Masson, Paris, 1983.

[Bre04] Claude Brezinski. Extrapolation algorithms for filtering series of functions, and treating the Gibbs phenomenon. *Numerical Algorithms*, 36(4):309–329, 2004.

[Bus06] V. Buslaev. On the Fabry ratio theorem for orthogonal series. *Proceedings of the Steklov Institute of Mathematics*, 253:8–21, 2006.

[Bus09] V. I. Buslaev. An analogue of Fabry's theorem for generalized Padé approximants. *Sbornik: Mathematics*, 200:981–1050, August 2009.

[Cab94] Ch. Cabos. A preconditioning of the tau operator for ordinary differential equations. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 74(11):521–532, 1994.

[Car26] T. Carleman. *Les fonctions quasi-analytiques*. Gauthier-Villars, 1926.

[CH89] R. Courant and D. Hilbert. *Methods of Mathematical Physics: Volume I*. Wiley, New York, 1989.

[Che66] E. W. Cheney. *Introduction to approximation theory*. McGraw-Hill New York, 1966.

[Chi73] J. S. R. Chisholm. Rational approximants defined from double power series. *Math. Comp.*, 27(124):841–848, October 1973.

[CHQZ07] C.G. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral Methods: Fundamentals in single domains*. Springer-Verlag New York, Inc., 2007.

[CL74] C.W. Clenshaw and K. Lord. *Rational approximation from Chebyshev series*. In: Studies in Numerical Analysis. Academic Press, London., 1974.

[CM74] J. S. R. Chisholm and J. McEwan. Rational approximants defined from power series in N variables. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 336:421–452, 1974.

[Cuy84] A. Cuyt. *Padé Approximants for Operators: Theory and Applications*, volume 1065 of *Lectures Notes in Mathematics*. Springer, Berlin, 1984.

[Cuy99] A. Cuyt. How well can the concept of Padé approximant be generalized to the multivariate case? *J. Comput. Appl. Math.*, 105:25–50, May 1999.

[D.64] Elliot D. The evaluation and estimation of the coefficients in the Chebyshev series expansion of a function. *Math. Comput.*, 18(18):274 – 284, 1964.

[Die57] P. Dienes. *The Taylor series*. Dover, Dover, New York, second edition, 1957.

[dMdB02] R. de Montessus de Ballore. Sur les fractions continues algébriques. *Bull. Soc. Math. France*, 30:28 – 36, 1902.

[DR84] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Academic Press, Academic Press: London, second edition, 1984.

[Fab96] E. Fabry. Sur les points singuliers d'une fonction donnée par son développement en série et l'impossibilité du prolongement analytique dans des cas très généraux. *Ann. Ec. Norm. Sup.*, 13(3):367–399, 1896.

[Fle73] J. Fleischer. Nonlinear Padé approximants for Legendre series. *J. Math. and Physics*, 14(2):246–248, 1973.

[Fro69] M. Froissart. Approximation de Padé: application à la physique des particules élémentaires. Technical Report 25, CNRS, Strasbourg, 1969.

[FS66] B. A. Finlayson and L. E. Scriven. The method of weighted residuals – a review. *Applied Mechanics Reviews*, 19(9):735–748, 1966.

[Fun92] D. Funaro. *Polynomial Approximation of Differential Equations, Lecture Notes in Physics, Vol 8*. Springer-Verlag, Heidelberg:, 1992.

[GGT13] P. Gonnet, S. Güttel, and L. N. Trefethen. Robust Padé Approximation via SVD. *SIAM Review*, 55(1):101–117, 2013.

[GH66] W. B. Gragg and A. S. Householder. On a theorem of König. *Numerische Mathematik*, 8(5):465–468, August 1966.

[GN73] J.L Gammel and J Nuttall. Convergence of Padé approximants to quasianalytic functions beyond natural boundaries. *J. Math. Comp. App.*, 43(3):694 – 696, 1973.

[GR07] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products.* Elsevier/Academic Press, Amsterdam., seventh edition, 2007.

[GTV87] J. Gilewicz and B. Truong-Van. Froissart doublets in the Padé approximation and noise. Technical Report CPT-2014. M-CPT-2014, CNRS Marseille. Cent. Phys. Théor., Marseille, Jun 1987.

[Gui97] P. Guillaume. Nested multivariate Padé approximants. *J. Math. Comp. App.*, 82(1):149 – 158, 1997.

[Gui98] P. Guillaume. Convergence of the nested multivariate Padé approximants. *Journal of Approximation Theory*, 94(3):455 – 466, 1998.

[Hol69] T.J. Holdeman, Jr. A method for the approximation of functions defined by formal series expansions in orthogonal polynomials. *Math. Comput.*, (23):275–287, 1969.

[Kah93] J.P. Kahane. *Some Random Series of Functions*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1993.

[Kön84] J. König. Über eine eigenschaft der potenzreihen. *Math. Ann*, 23:447–449, 1884.

[Kuf85] A. Kufner. *Weighted Sobolev spaces*. John Wiley and Sons, Inc. New York, 1985.

[Luk69] Yudell L. Luke. *The Special Functions and their Approximations*, volume 1. Academic Press, 1969.

[Mat01] A. C. Matos. Recursive computation of Padé-Legendre approximants and some acceleration properties. *Numerische Mathematik*, 89:535–560, 2001.

[Mat03] J. Matos. *Algoritmos de cálculo dos aproximantes de Frobenius-Padé e generalizações*. PhD thesis, Faculdade de Ciências da Universidade do Porto, 2003.

[Mat07] A. C. Matos. Multivariate Frobenius-Padé; approximants: Properties and algorithms. *J. Comput. Appl. Math.*, 202:548–572, 2007.

[MMR14] J.C. Matos, J. Matos, and M. J. Rodrigues. On the Localization of Zeros and Poles of Chebyshev-Padé Approximants from Perturbed Functions. In *Computational Science and Its Applications ICCSA 2014*, volume 8584 of *Lecture Notes in Computer Science*, pages 481–492. Springer International Publishing, 2014.

[MRMC] J. Matos, M.J. Rodrigues, J.C. Matos, and M. Cruz. Avoiding similarity transformations in the operational tau method. Submitted to BIT Numerical Mathematics (2014).

[MRV04] J. Matos, M. J. Rodrigues, and P. B. Vasconcelos. New implementation of the tau method for PDEs. *J. Comput. Appl. Math.*, 164-165:555–567, 2004.

[Neč62] J. Nečas. Sur une méthode pour resoudre les equations aux derivées partielles du type elliptique, voisine de la variationnelle. *Ann. Sc. Norm. Sup. Pisa*, (16):305–326, 1962.

[Nut70] J. Nuttall. The convergence of Padé approximants of meromorphic functions. *Journal of Mathematical Analysis and Applications*, (31):147–153, 1970.

[OS80] Eduardo L. Ortiz and H. Samara. A new operational approach to the numerical solution of differential equations in terms of polynomials. *in Innovative Numerical Analysis for the Engineering Sciences (R. Shaw and W. Pilkey, Eds.)*, 27:643–652, 1980. The University Press of Virginia.

[OS81] E.L. Ortiz and H. Samara. An operational approach to the tau method for the numerical solution of non-linear differential equations. *Computing*, 27(1):15–25, 1981.

[Pas84] S. Paszkowski. Polynômes et séries de Tchebichev. Technical report, Univ. Lille 1, 1984.

[Pey02] R. Peyret. *Spectral Methods for Incompressible Viscous Flow*, volume 148 of *Applied Mathematical Sciences*. Springer, New York, 2002.

[Pom73] Ch. Pommerenke. Padé approximants and convergence in capacity. *J. Math. Comp. App.*, 41(3):775 – 780, 1973.

[Riv74] T.J. Rivlin. *The Chebyshev polynomials*. Pure and applied mathematics. Wiley, 1974.

[RP89] H. G. Roos and E. Pfeifer. A convergence result for the tau method. *Computing*, 42:81–84, 1989.

[Sch66] L. Schwarz. *Théorie des Distributions*. Hermann, Paris, 1966.

[Sid03] A. Sidi. *Practical Extrapolation Methods: Theory and Applications*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2003.

[Sta98] H. Stahl. Spurious poles in Padé approximation. *J. Comput. App. Math.*, 99(1):511 – 527, 1998. Proceeding of the {VIIIth} Symposium on Orthogonal Polynomials and Their Application.

[Sue85] S. P. Suetin. On an inverse problem for the m th row of the Padé table. *Mathematics of the USSR-Sbornik*, 52(1):231, 1985.

[Sze39] G. Szegö. *Orthogonal polynomials*. American Mathematical Society, Providence, 4th ed. edition, 1939.

[Tim63] A. F. Timan. *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, Oxford, 1963.

[VGP81] V. V. Vavilov, G.Lopes, and V. A. Prohorov. On an inverse problem for the rows of a Padé table. *Mathematics of the USSR-Sbornik*, 38(1):109, 1981.

[ZC67] O. C. Zienkiewicz and Y. K Cheung. *The Finite Method in Strutural and Continuum Mechanics*. McGraw-Hill, London, 1967.